Sebastián Pastor Ferrari

Data Mining – Assignment 01

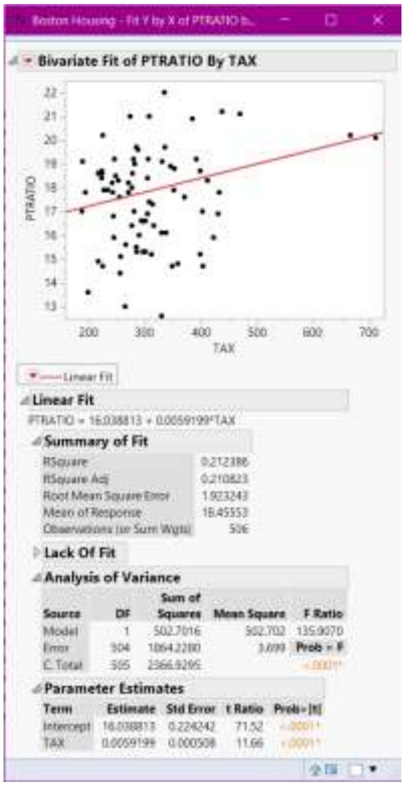
02.03.2020

Question 1 – Data Exploration :

a) **Generate a random sample from the data set that contains 80% of the rows (use the seed = 4279), and answer the following questions with respect to this sample.**
Done.

b) **How many rows are in this data set? How many columns? What do the rows and columns represent?**
There are 506 rows, which each represents a different household in the Boston region. The columns are all different characteristics for each household.

c) **Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.**


Boston Housing - Fit Y by X of PTRATIO b...
Bivariate Fit of PTRATIO By TAX
PTRATIO
22
21
20
19
18
17
16
15
14
13
200
300
400
500
600
700
TAX
Linear Fit

Linear Fit

PTRATIO = 16.038813 + 0.0059199*TAX

Summary of Fit

RSquare	0.212386
RSquare Adj	0.210823
Root Mean Square Error	1.923243
Mean of Response	18.45553
Observations (or Sum Wgts)	506

Lack Of Fit

Analysis of Variance

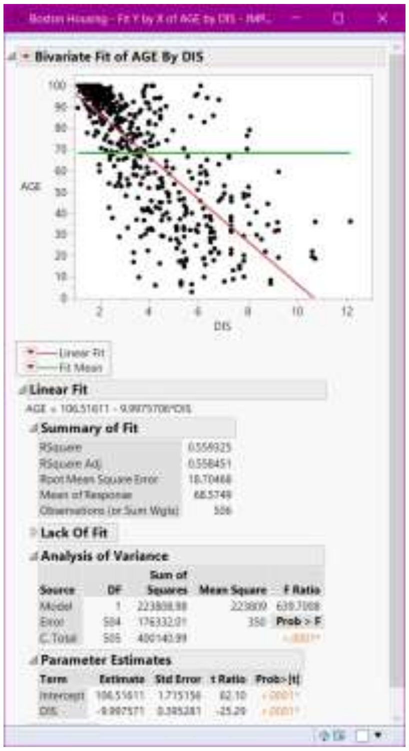
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	502.7016	502.702	135.9070
Error	504	1864.2280	3.699	Prob > F
C. Total	505	2366.9295		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	16.038813	0.224242	71.52	<.0001*
TAX	0.0059199	0.000508	11.66	<.0001*

i. x(property tax) by y(pupils per teacher). Property tax is not a good indicator of pupils per teacher. I would have expected that towns with more property tax revenue, to have larger investments in schools (by number of teachers). The opposite is true however, as property taxes go up, the number of pupils-to-teacher goes up as well.

ii. x(weighted distance to five employment centers) by y(proportion of households built before 1940). I expected newer houses to be further away from dense employment centers due to the growth of suburban towns. My assessment was correct in this case.


Boston Housing - Fit Y by X of AGE by DIS - JMP...
Bivariate Fit of AGE By DIS
AGE
100
90
80
70
60
50
40
30
20
10
0
2
4
6
8
10
12
DIS
Linear Fit
Fit Mean

Linear Fit

AGE = 106.51611 - 9.9975706*DIS

Summary of Fit

RSquare	0.559325
RSquare Adj	0.558451
Root Mean Square Error	18.70468
Mean of Response	68.5749
Observations (or Sum Wgts)	506

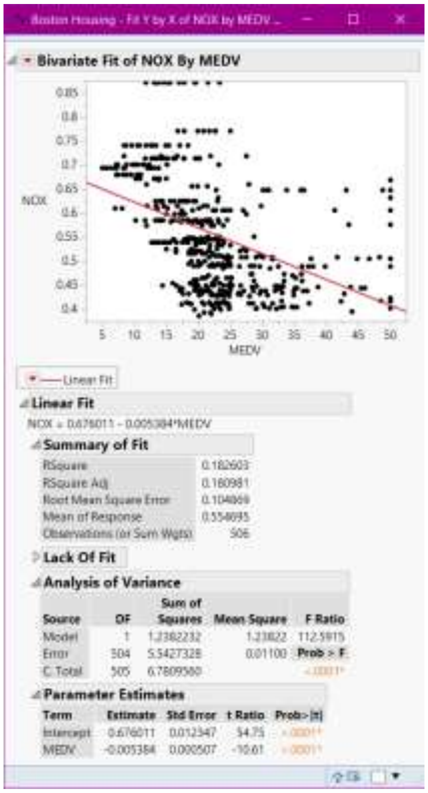
Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	223808.98	223809	639.7088
Error	504	176332.01	350	Prob > F
C. Total	505	400140.99		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	106.51611	1.715156	62.10	<.0001*
DIS	-9.997571	0.395281	-25.29	<.0001*



Bivariate Fit of NOX By MEDV

Linear Fit

NOX = 0.676011 - 0.005384*MEDV

Summary of Fit

RSquare	0.182603
RSquare Adj	0.180981
Root Mean Square Error	0.104869
Mean of Response	0.554695
Observations (or Sum Wgts)	506

Lack Of Fit

Analysis of Variance

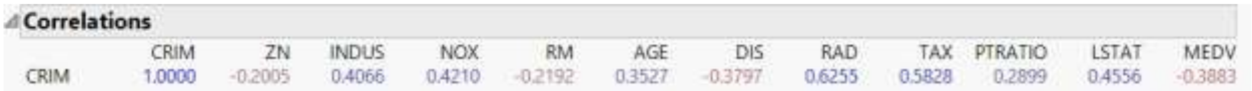
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1.2382232	1.23822	112.5915
Error	504	5.5427328	0.01100	Prob > F
C. Total	505	6.7809560		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.676011	0.012347	54.75	<.0001*
MEDV	-0.005384	0.000507	-10.61	<.0001*

iii. x(Median Value of house) by y(Nitric Oxide Concentration). I expected the concentration of NO in the environment to go down in more expensive towns. However, there seems to be no pattern. Either the people of Boston are not aware of pollution or they don't seem to care about it, indicated by the housing market data we have.

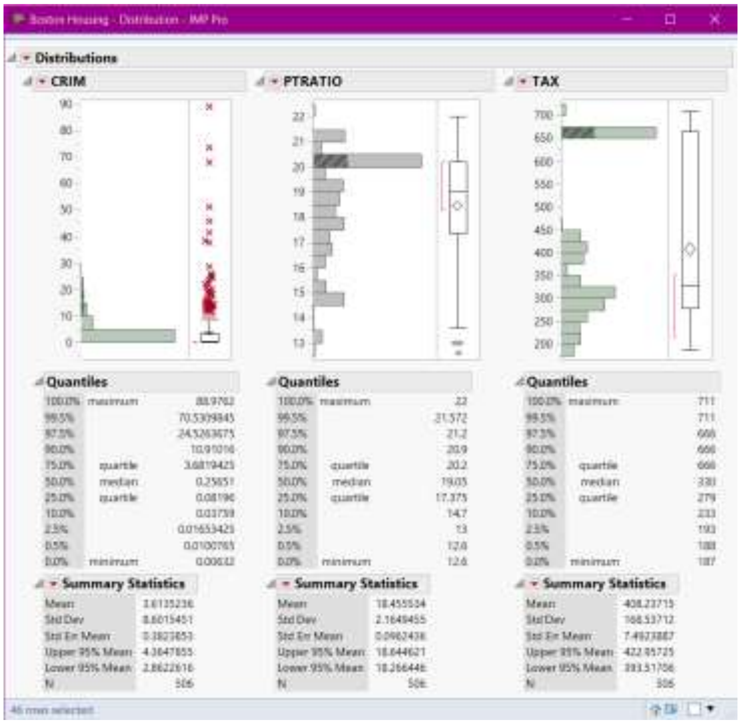
d) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.



Correlations

	CRIM	ZN	INDUS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	LSTAT	MEDV
CRIM	1.0000	-0.2005	0.4066	0.4210	-0.2192	0.3527	-0.3797	0.6255	0.5828	0.2899	0.4556	-0.3883

Per Capita Crime rate is associated with both RAD(Index of Accessibility to highways), and TAX(Full-Value Property Tax Rate). CRIM/RAD makes sense due to criminals wanting good ways of getting away and probably perform crimes near gas stations, fast foods, etc. Full property tax value is surprising, the reasonable explanation for this relationship could be that homes in the metropolitan have higher taxes, and crime takes more place in urban regions.



Distributions

CRIM

Quantiles

100.0%	maximum	88.9762
99.5%		70.5309845
97.5%		24.5263675
90.0%		10.91016
75.0%	quartile	3.6819425
50.0%	median	0.25651
25.0%	quartile	0.08196
10.0%		0.03739
2.5%		0.01653425
0.5%		0.0100765
0.0%	minimum	0.00632

Summary Statistics

Mean	3.6135236
Std Dev	8.6015451
Std Err Mean	0.3823853
Upper 95% Mean	4.3647855
Lower 95% Mean	2.8622616
N	506

PTRATIO

Quantiles

100.0%	maximum	22
99.5%		21.572
97.5%		21.2
90.0%		20.9
75.0%	quartile	20.2
50.0%	median	19.05
25.0%	quartile	17.375
10.0%		14.7
2.5%		13
0.5%		12.6
0.0%	minimum	12.6

Summary Statistics

Mean	18.455534
Std Dev	2.1649455
Std Err Mean	0.0962436
Upper 95% Mean	18.644621
Lower 95% Mean	18.266446
N	506

TAX

Quantiles

100.0%	maximum	711
99.5%		711
97.5%		666
90.0%		666
75.0%	quartile	666
50.0%	median	330
25.0%	quartile	279
10.0%		233
2.5%		193
0.5%		188
0.0%	minimum	187

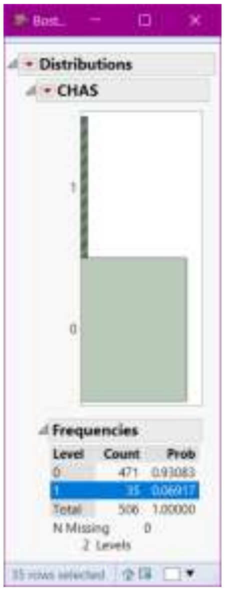
Summary Statistics

Mean	408.23715
Std Dev	168.53712
Std Err Mean	7.4923887
Upper 95% Mean	422.95725
Lower 95% Mean	393.51706
N	506

46 rows selected

e) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

There is a range of suburbs that are not marginally, but greatly away from the normal range, particularly in Crime Rate. The mean value for crime rate is 3.6, yet a great deal of towns go from 10 all the way to 90. These towns have particularly high pupil to teacher ratio and high property tax ratios. The Pupil to teacher Ratio has a mean of 18.45, and Tax has a ratio of 408.23.



Frequencies

Level	Count	Prob
0	471	0.93083
1	35	0.06917
Total	506	1.00000

N Missing 0
2 Levels

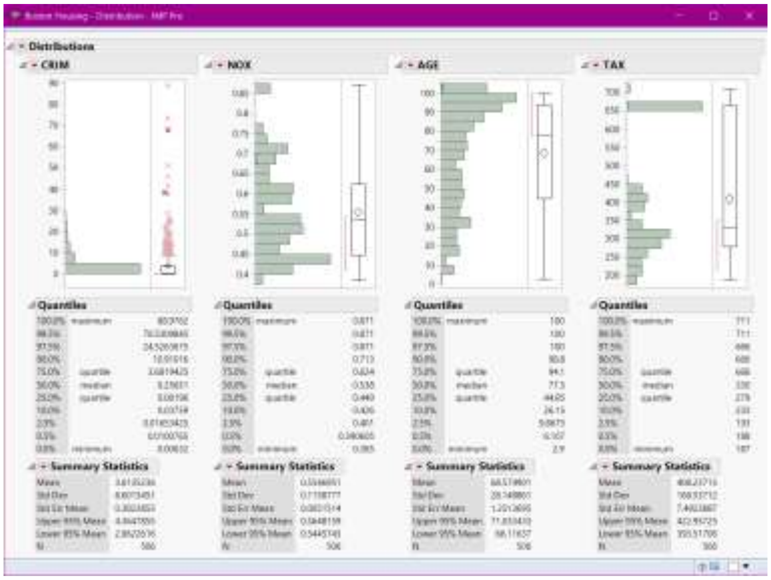
35 rows selected

f) How many of the suburbs in this data set bound the Charles river?

35 out of the 506 towns are set bound the Charles River.

g) What is the median pupil-teacher ratio among the towns in this data set?

As mentioned above, the mean of Pupil to Teacher ratio is 18.45.



CRIM Quantiles

100.0%	maximum	88.9762
99.5%		76.2[illegible]
97.5%		24.5260815
90.0%		10.91016
75.0%	quartile	3.6819425
50.0%	median	0.25651
25.0%	quartile	0.08196
10.0%		0.03759
2.5%		0.01853425
0.5%		0.01100765
0.0%	minimum	0.00632

CRIM Summary Statistics

Mean	3.6135236
Std Dev	8.6015451
Std Err Mean	0.3823853
Upper 95% Mean	4.3647055
Lower 95% Mean	2.8623616
N	506

NOX Quantiles

100.0%	maximum	0.871
99.5%		0.871
97.5%		0.871
90.0%		0.713
75.0%	quartile	0.624
50.0%	median	0.538
25.0%	quartile	0.449
10.0%		0.426
2.5%		0.401
0.5%		0.390605
0.0%	minimum	0.385

NOX Summary Statistics

Mean	0.5546951
Std Dev	0.1158777
Std Err Mean	0.0051514
Upper 95% Mean	0.5648159
Lower 95% Mean	0.5445743
N	506

AGE Quantiles

100.0%	maximum	100
99.5%		100
97.5%		100
90.0%		98.8
75.0%	quartile	94.1
50.0%	median	77.5
25.0%	quartile	44.85
10.0%		26.15
2.5%		9.0875
0.5%		6.107
0.0%	minimum	2.9

AGE Summary Statistics

Mean	68.574901
Std Dev	28.148861
Std Err Mean	1.2513695
Upper 95% Mean	71.033403
Lower 95% Mean	66.11637
N	506

TAX Quantiles

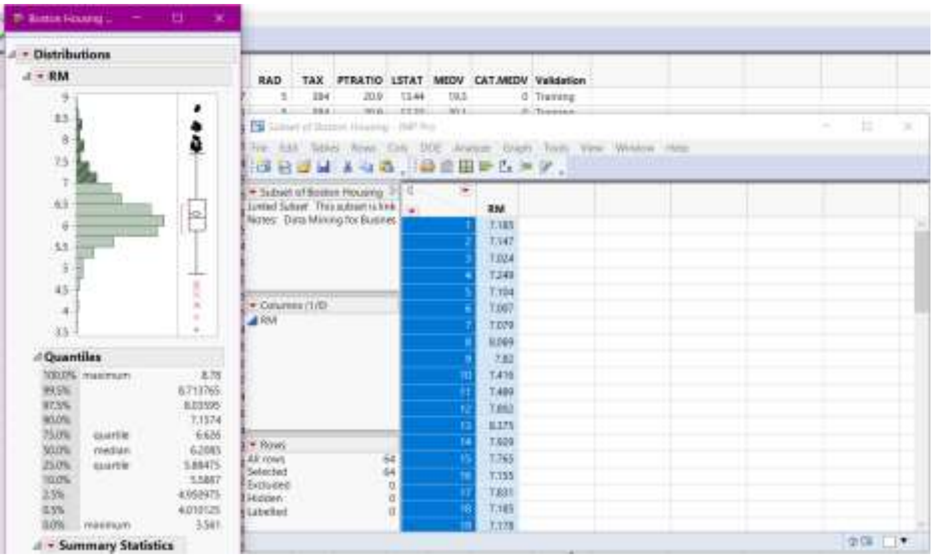
100.0%	maximum	711
99.5%		711
97.5%		666
90.0%		666
75.0%	quartile	666
50.0%	median	330
25.0%	quartile	279
10.0%		233
2.5%		193
0.5%		188
0.0%	minimum	187

TAX Summary Statistics

Mean	408.23715
Std Dev	168.53712
Std Err Mean	7.4922887
Upper 95% Mean	422.95725
Lower 95% Mean	393.51705
N	506

h) Which suburb of Boston has lowest median value of owner occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

Rows 309 and 406 both have the minimum value (5). They are characterized by high Age, Tax, Crime Rate, and Nitrous Oxide levels.


Boston Housing
Distributions
RM
9
8.5
8
7.5
7
6.5
6
5.5
5
4.5
4
3.5
Subset of Boston Housing - JMP Pro
File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help
Subset of Boston Housing
Limited Subset: This subset is link
Notes: Data Mining for Busines
Columns (1/0)
RM
Summary Statistics

RAD	TAX	PTRATIO	LSTAT	MEDV	CAT.MEDV	Validation
5	384	20.9	13.44	19.5	0	Training

Quantiles

100.0%	maximum	8.78
99.5%		8.713765
97.5%		8.03595
90.0%		7.1574
75.0%	quartile	6.626
50.0%	median	6.2085
25.0%	quartile	5.88475
10.0%		5.5887
2.5%		4.950975
0.5%		4.010125
0.0%	minimum	3.561

	RM
1	7.185
2	7.147
3	7.024
4	7.249
5	7.104
6	7.007
7	7.079
8	8.069
9	7.82
10	7.416
11	7.489
12	7.802
13	8.375
14	7.929
15	7.765
16	7.155
17	7.831
18	7.185
19	7.178

Rows

All rows	64
Selected	64
Excluded	0
Hidden	0
Labelled	0

i) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

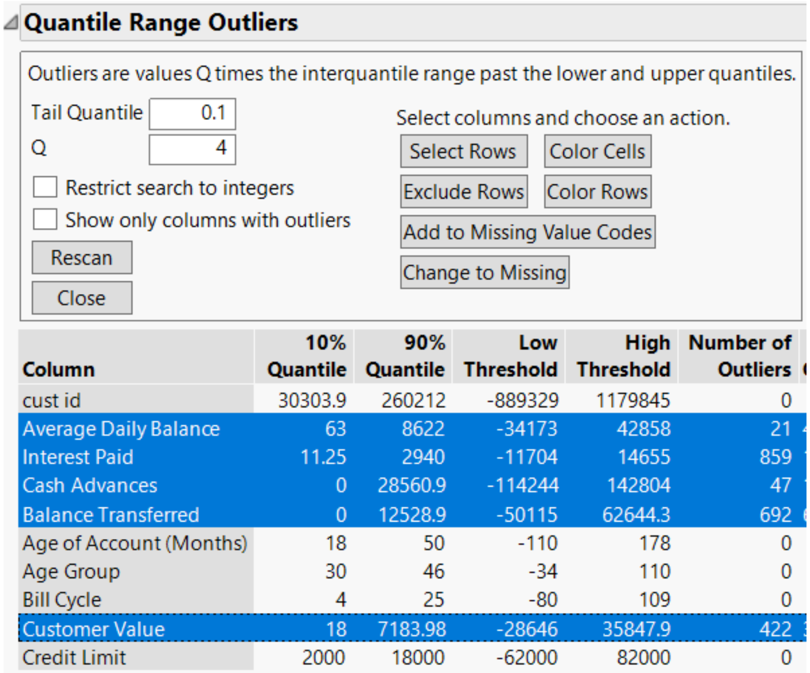
64 suburbs average more than 7 rooms per dwelling.

13 average more than 8 rooms per dwelling. These have high Median Values, Large Size, Large Age Values, and low Crime Rates, low Pupil to Teacher ratio, as well as low % Of Low Status Population.

Question 2 – Data Processing:

a) **Summarize the variables in the data set – which variables are continuous, nominal, ordinal?**
Continous: Average Daily Balance, Interest Paid, Cash Advances, Balance Transferred, Age Group, Customer Value
Nominal: cust id, Credit Limit, Marital Status, Occupation Group, Customer Type, Gender
Ordinal: LTV Group, Age of Account (Months), Bill Cycle,

b) **Identify if there any outliers.**
outlier report detailed below.



Quantile Range Outliers

Outliers are values Q times the interquantile range past the lower and upper quantiles.

Tail Quantile 0.1
Q 4

□ Restrict search to integers
□ Show only columns with outliers

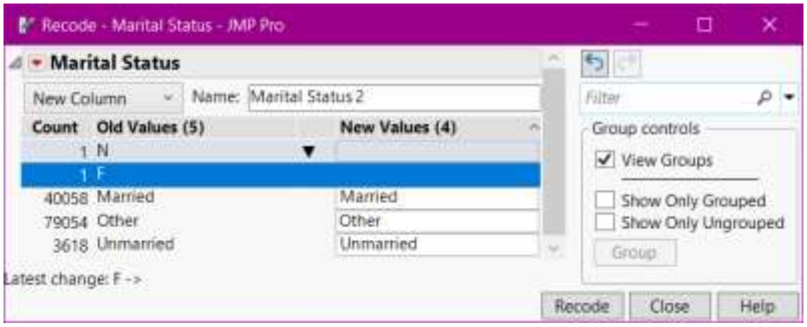
Rescan
Close

Select columns and choose an action.

Column	10% Quantile	90% Quantile	Low Threshold	High Threshold	Number of Outliers	(
cust id	30303.9	260212	-889329	1179845	0	
Average Daily Balance	63	8622	-34173	42858	21	
Interest Paid	11.25	2940	-11704	14655	859	
Cash Advances	0	28560.9	-114244	142804	47	
Balance Transferred	0	12528.9	-50115	62644.3	692	
Age of Account (Months)	18	50	-110	178	0	
Age Group	30	46	-34	110	0	
Bill Cycle	4	25	-80	109	0	
Customer Value	18	7183.98	-28646	35847.9	422	
Credit Limit	2000	18000	-62000	82000	0	

c) Examine Marital Status, LTV Group, and Gender. Are there any typographical errors? If so, correct them.

For gender, instead of "____", I recoded it to "n/a" for easier reference. In marital status, there were two instances ("N", and "F"). I recoded both to "_____", because it is safer than assuming what they mean, and there is enough data to afford it.



Recode - Marital Status - JMP Pro

Marital Status

New Column Name: Marital Status 2

Count	Old Values (5)	New Values (4)
1	N	
1	F	
40058	Married	Married
79054	Other	Other
3618	Unmarried	Unmarried

Latest change: F ->

Group controls

☑ View Groups

□ Show Only Grouped

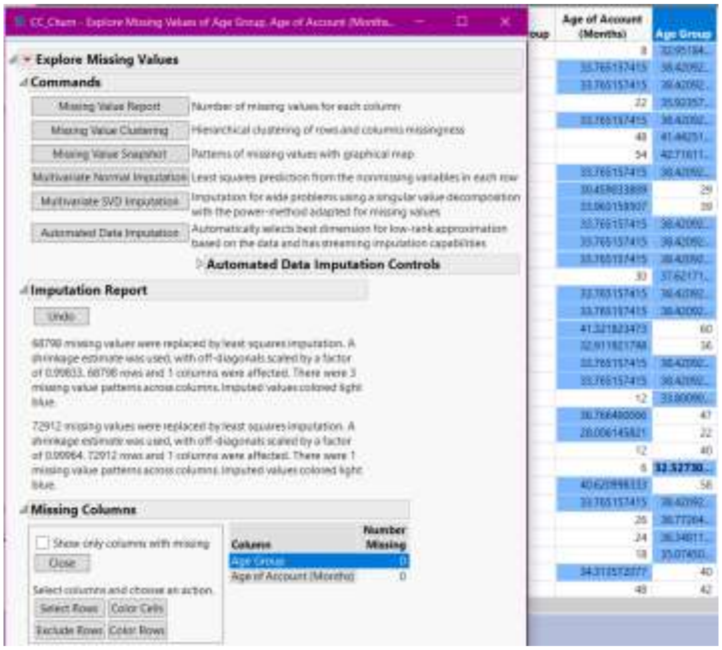
□ Show Only Ungrouped

d) Examine Gender. Are there any missing values? If so, how many?

Gender is missing 59,234 entries.

e) Examine Age of Account (Months) and Age Group. Are there any missing values? If so, how many? Impute them

Age of Accounts is missing 68, 798 values, while Age Group is missing 72,912 values. I made up for this with Multivariate Normal Computation.



CC_Churn - Explore Missing Values of Age Group, Age of Account (Months...

Explore Missing Values

Commands

Missing Value Report	Number of missing values for each column
Missing Value Clustering	Hierarchical clustering of rows and columns missingness
Missing Value Snapshot	Patterns of missing values with graphical map
Multivariate Normal Imputation	Least squares prediction from the nonmissing variables in each row
Multivariate SVD Imputation	Imputation for wide problems using a singular value decomposition with the power-method adapted for missing values
Automated Data Imputation	Automatically selects best dimension for low-rank approximation based on the data and has streaming imputation capabilities

Automated Data Imputation Controls

Imputation Report

68798 missing values were replaced by least squares imputation. A shrinkage estimate was used, with off-diagonals scaled by a factor of 0.99833. 68798 rows and 1 columns were affected. There were 3 missing value patterns across columns. Imputed values colored light blue.

72912 missing values were replaced by least squares imputation. A shrinkage estimate was used, with off-diagonals scaled by a factor of 0.99964. 72912 rows and 1 columns were affected. There were 1 missing value patterns across columns. Imputed values colored light blue.

Missing Columns

□ Show only columns with missing

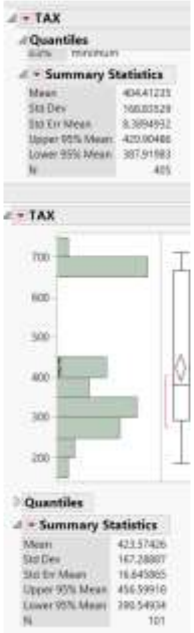
Select columns and choose an action.

Column	Number Missing
Age Group	0
Age of Account (Months)	0

Age of Account (Months)	Age Group
8	32.95184...
33.765157415	38.42092...
33.765157415	38.42092...
22	35.92357...
33.765157415	38.42092...
48	41.46251...
54	42.71611...
33.765157415	38.42092...
30.458033889	29
33.060158307	39
33.765157415	38.42092...
33.765157415	38.42092...
33.765157415	38.42092...
30	37.62171...
33.765157415	38.42092...
33.765157415	38.42092...
41.321823473	60
32.911821798	36
33.765157415	38.42092...
33.765157415	38.42092...
12	33.80090...
30.766400000	47
28.006145821	22
12	40
6	32.52730...
40.620998133	58
33.765157415	38.42092...
26	38.77204...
24	36.34811...
18	35.07450...
34.313572077	40
48	42

Question 3 – Prediction Using Regression

a. **Use the subset that you created in Question 1 as the training data set, and the rest of the data as the validation data set. Compare the variable summaries across the training and validation data sets. Do you have any concerns regarding the partition?**



TAX

Quantiles

Summary Statistics

Mean	404.41235
Std Dev	168.03528
Std Err Mean	8.3894932
Upper 95% Mean	420.90488
Lower 95% Mean	387.91983
N	405

TAX

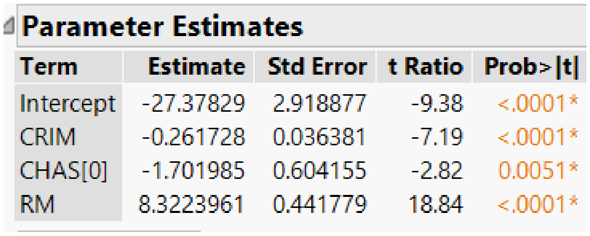
Quantiles

Summary Statistics

Mean	423.57426
Std Dev	167.28887
Std Err Mean	16.645865
Upper 95% Mean	456.59918
Lower 95% Mean	390.54934
N	101

Tax has the most significantly different means, and the most spread (Std. Deviation). The rest of the data does not have this amount of spread.

b. **Fit a multiple linear regression model to the median house price (MEDV) as a function of CRIM, CHAS, and RM using the training data Write the equation for predicting the median house price from the predictors in the model.**



Parameter Estimates

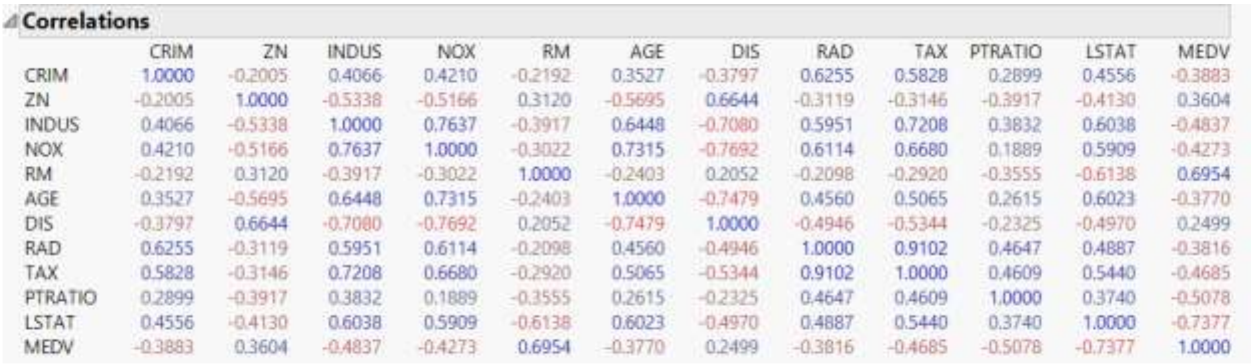
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-27.37829	2.918877	-9.38	<.0001*
CRIM	-0.261728	0.036381	-7.19	<.0001*
CHAS[0]	-1.701985	0.604155	-2.82	0.0051*
RM	8.3223961	0.441779	18.84	<.0001*

MEDV = -27.38 -0.26(CRIM) -1.7(CHAS: 0/1) + 8.3(RM)

c. **Using the estimated regression model, what median house price is predicted for a tract in the Boston area that does not bound the Charles River, has a crime rate of 0.1, and where the average number of rooms per house is 6? What is the prediction error?**
$23,394

d. What is the RMSE on the validation data?
6.3

d. Consider the 12 predictors:



Correlations

	CRIM	ZN	INDUS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	LSTAT	MEDV
CRIM	1.0000	-0.2005	0.4066	0.4210	-0.2192	0.3527	-0.3797	0.6255	0.5828	0.2899	0.4556	-0.3883
ZN	-0.2005	1.0000	-0.5338	-0.5166	0.3120	-0.5695	0.6644	-0.3119	-0.3146	-0.3917	-0.4130	0.3604
INDUS	0.4066	-0.5338	1.0000	0.7637	-0.3917	0.6448	-0.7080	0.5951	0.7208	0.3832	0.6038	-0.4837
NOX	0.4210	-0.5166	0.7637	1.0000	-0.3022	0.7315	-0.7692	0.6114	0.6680	0.1889	0.5909	-0.4273
RM	-0.2192	0.3120	-0.3917	-0.3022	1.0000	-0.2403	0.2052	-0.2098	-0.2920	-0.3555	-0.6138	0.6954
AGE	0.3527	-0.5695	0.6448	0.7315	-0.2403	1.0000	-0.7479	0.4560	0.5065	0.2615	0.6023	-0.3770
DIS	-0.3797	0.6644	-0.7080	-0.7692	0.2052	-0.7479	1.0000	-0.4946	-0.5344	-0.2325	-0.4970	0.2499
RAD	0.6255	-0.3119	0.5951	0.6114	-0.2098	0.4560	-0.4946	1.0000	0.9102	0.4647	0.4887	-0.3816
TAX	0.5828	-0.3146	0.7208	0.6680	-0.2920	0.5065	-0.5344	0.9102	1.0000	0.4609	0.5440	-0.4685
PTRATIO	0.2899	-0.3917	0.3832	0.1889	-0.3555	0.2615	-0.2325	0.4647	0.4609	1.0000	0.3740	-0.5078
LSTAT	0.4556	-0.4130	0.6038	0.5909	-0.6138	0.6023	-0.4970	0.4887	0.5440	0.3740	1.0000	-0.7377
MEDV	-0.3883	0.3604	-0.4837	-0.4273	0.6954	-0.3770	0.2499	-0.3816	-0.4685	-0.5078	-0.7377	1.0000

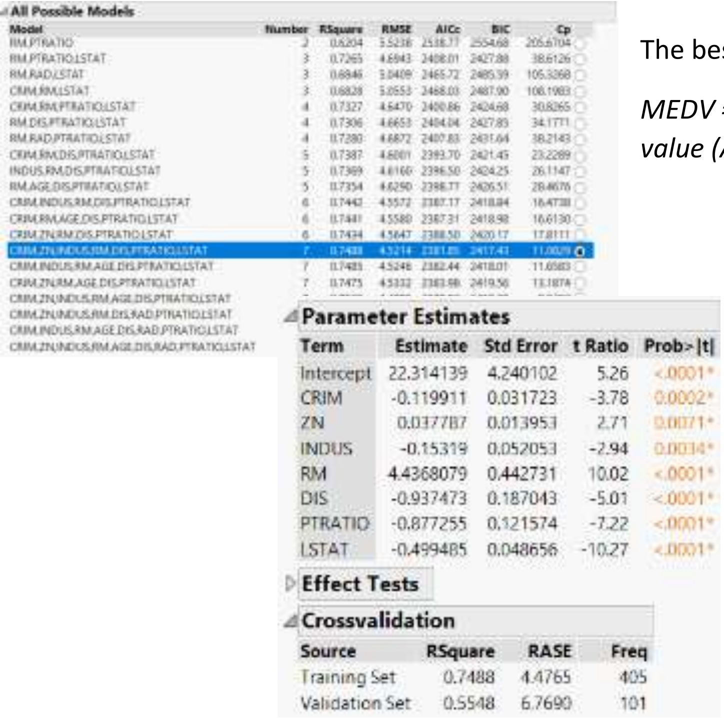
i. **Which predictors are likely to be measuring the same thing among the entire set of predictors? Discuss the relationships among INDUS, NOX, and TAX.**

RAD-TAX which have a .91 correlation coefficient. INDUS-NOX have a high coefficient (0.763), as well as INUDS-TAX (0.721), as well as NOX-TAX (0.668).

ii. **Compute the correlation table for the numerical predictors and search for highly correlated pairs. These have potential redundancy and can cause multicollinearity. Choose which ones to remove based on this table.**

I will remove TAX and NOX, given they are highly correlated with each other and with other predictors.

iii. **Use an exhaustive search (All Possible Models) to reduce the remaining predictors as follows: First, choose the top three models. Then run each of these models and compare their predictive accuracy for the validation set. Compare RMSE, Cp, $AICc$, and Validation RSquare. Finally, describe the best model.**



All Possible Models

Model	Number	RSquare	RMSE	AICc	BIC	Cp
RM,PTRATIO	2	0.6204	5.5236	2538.77	2554.68	206.6704 ○
RM,PTRATIO,LSTAT	3	0.7265	4.6943	2408.01	2427.88	38.6126 ○
RM,RAD,LSTAT	3	0.6846	5.0409	2465.72	2485.39	105.3268 ○
CRIM,RM,LSTAT	3	0.6828	5.0553	2468.03	2487.90	108.1983 ○
CRIM,RM,PTRATIO,LSTAT	4	0.7327	4.6470	2400.86	2424.68	30.8265 ○
RM,DIS,PTRATIO,LSTAT	4	0.7306	4.6653	2404.04	2427.83	34.1771 ○
RM,RAD,PTRATIO,LSTAT	4	0.7280	4.6872	2407.83	2431.64	38.2143 ○
CRIM,RM,DIS,PTRATIO,LSTAT	5	0.7387	4.6001	2393.70	2421.45	23.2289 ○
INDUS,RM,DIS,PTRATIO,LSTAT	5	0.7369	4.6160	2396.50	2424.25	26.1147 ○
RM,AGE,DIS,PTRATIO,LSTAT	5	0.7354	4.6290	2398.77	2426.51	28.4676 ○
CRIM,INDUS,RM,DIS,PTRATIO,LSTAT	6	0.7442	4.5572	2387.17	2418.84	16.4738 ○
CRIM,RM,AGE,DIS,PTRATIO,LSTAT	6	0.7441	4.5580	2387.31	2418.98	16.6130 ○
CRIM,ZN,RM,DIS,PTRATIO,LSTAT	6	0.7434	4.5647	2388.50	2420.17	17.8111 ○
CRIM,ZN,INDUS,RM,DIS,PTRATIO,LSTAT	7	0.7488	4.5214	2381.85	2417.43	11.0029 ●
CRIM,INDUS,RM,AGE,DIS,PTRATIO,LSTAT	7	0.7485	4.5246	2382.44	2418.01	11.6583 ○
CRIM,ZN,RM,AGE,DIS,PTRATIO,LSTAT	7	0.7475	4.5332	2383.98	2419.56	13.1874 ○
CRIM,ZN,INDUS,RM,AGE,DIS,PTRATIO,LSTAT						
CRIM,ZN,INDUS,RM,DIS,RAD,PTRATIO,LSTAT						
CRIM,INDUS,RM,AGE,DIS,RAD,PTRATIO,LSTAT						
CRIM,ZN,INDUS,RM,AGE,DIS,RAD,PTRATIO,LSTAT						

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	22.314139	4.240102	5.26	<.0001*
CRIM	-0.119911	0.031723	-3.78	0.0002*
ZN	0.037787	0.013953	2.71	0.0071*
INDUS	-0.15319	0.052053	-2.94	0.0034*
RM	4.4368079	0.442731	10.02	<.0001*
DIS	-0.937473	0.187043	-5.01	<.0001*
PTRATIO	-0.877255	0.121574	-7.22	<.0001*
LSTAT	-0.499485	0.048656	-10.27	<.0001*

Effect Tests

Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.7488	4.4765	405
Validation Set	0.5548	6.7690	101

The best model I found was:

MEDV = CRIM,ZN,INDUS,RM,DIS,PTRATIO,LSTAT, it has a predictive value (Adj. R^2) = 0.744.

These are the predictors with their respective estimates, as well as VIF for the Validation Dataset = 6.769.