Lab3

Subhashini R., Lei Yang, Ron Cordell

April 12, 2016

Lab 3

Part 1 - Modeling House Values

Exploratory Data Analysis

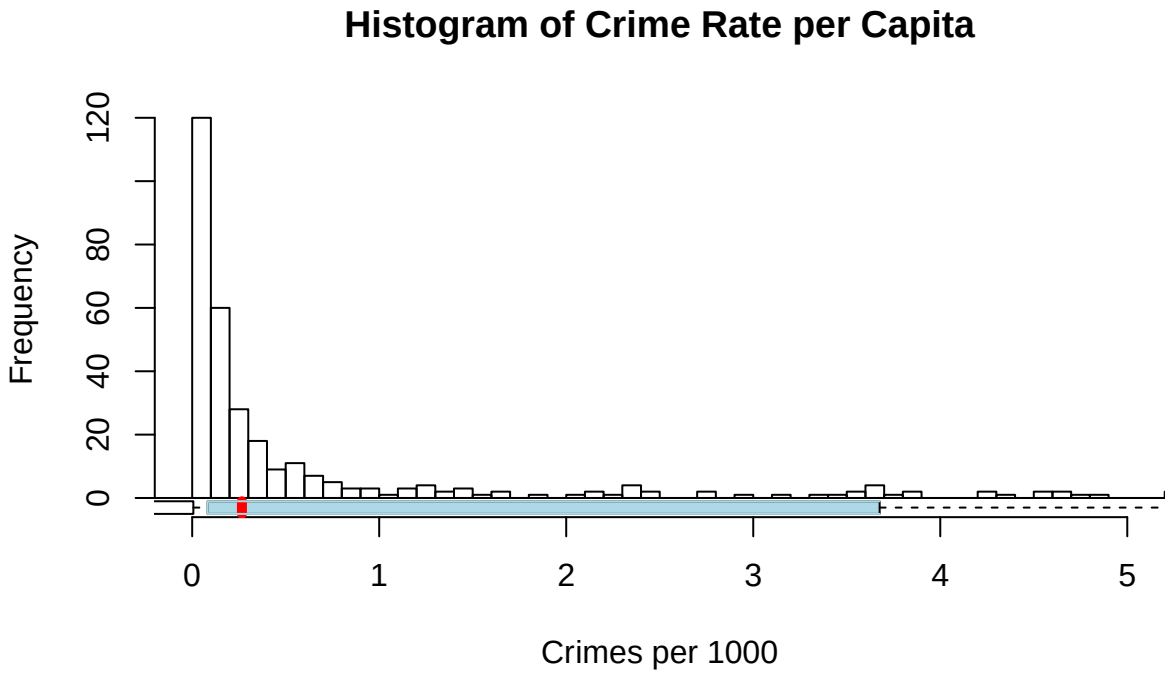
An examination of the provided data set reveals 11 variables of which *withWater* is binary and rest are continuous. There are no NA's in the data set, however the *distanceToHighway* variable appears to have a coding issue. We'll examine this variable in more detail a bit later, but first we summarize the variables in the following table.

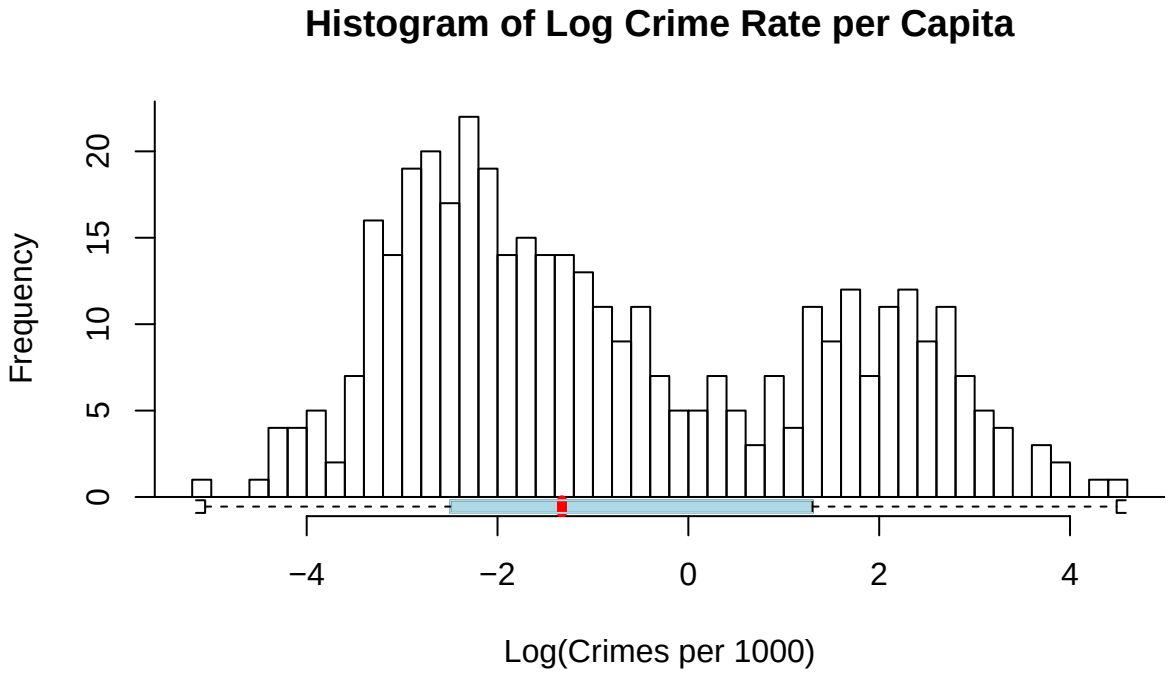
Table 1: Summary of Data

Statistic	N	Mean	St. Dev.	Min	Max
crimeRate_pc	400	3.763	8.872	0.006	88.976
nonRetailBusiness	400	0.112	0.070	0.007	0.277
withWater	400	0.068	0.251	0	1
ageHouse	400	68.932	27.977	2.900	100.000
distanceToCity	400	9.638	8.786	1.228	54.197
distanceToHighway	400	9.582	8.672	1	24
pupilTeacherRatio	400	21.391	2.168	15.600	25.000
pctLowIncome	400	15.795	9.341	2	49
homeValue	400	499,584.400	196,115.700	112,500	1,125,000
pollutionIndex	400	40.615	11.825	23.500	72.100
nBedRooms	400	4.266	0.719	1.561	6.780

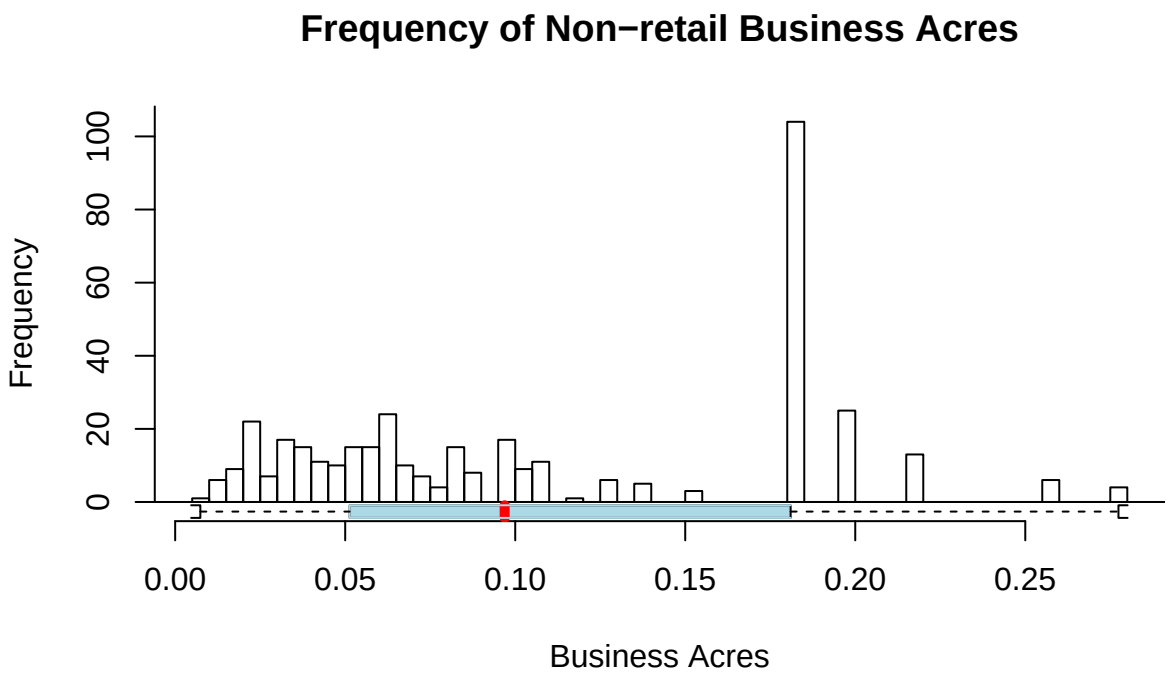
For the purposes of this analysis we categorize the variables *crimeRate_pc*, *nonRetailBusiness*, *withWater*, *distanceToCity*, *distanceToHighway*, *pollutionIndex*, *pupilTeacherRation* and *pctLowHousing* to be environmental variables. The variables *ageHouse* and *nBedRooms* are attributes of the house. The variable *homeValue* is the dependent variable we would like to explain in terms of primarily the environment variables but we will compare to explanations in terms of house attributes as well.

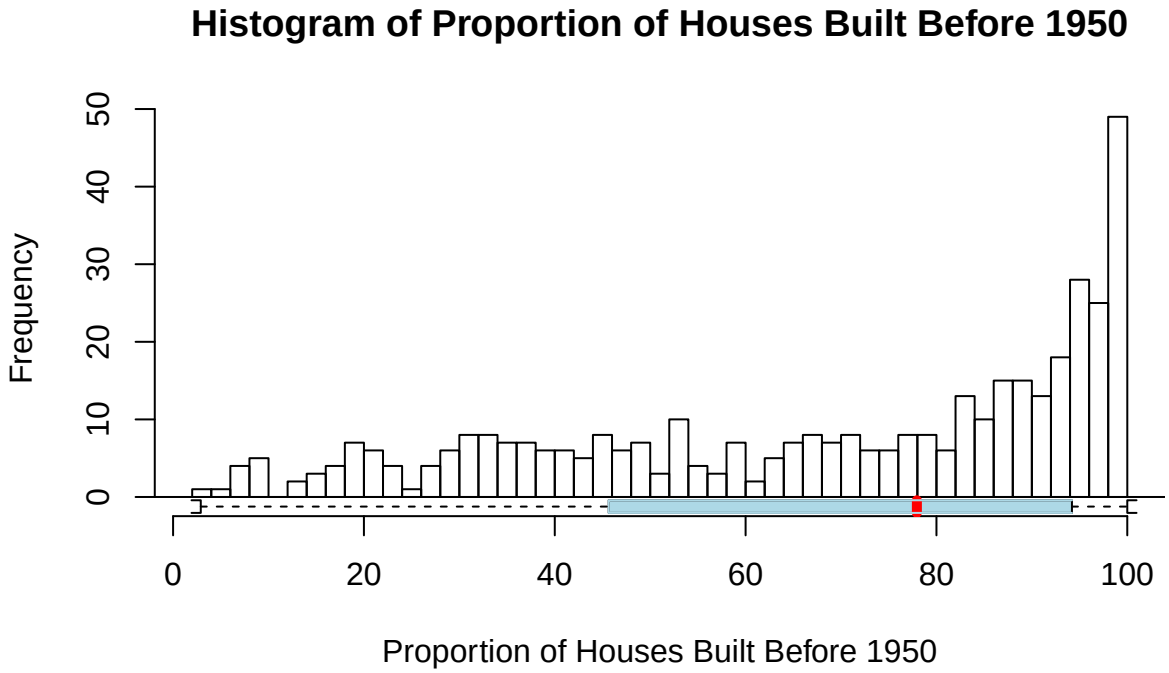
In the next several pages we examine the distribution of each of the variables and, where indicated, the distribution of the log(variable) as well.





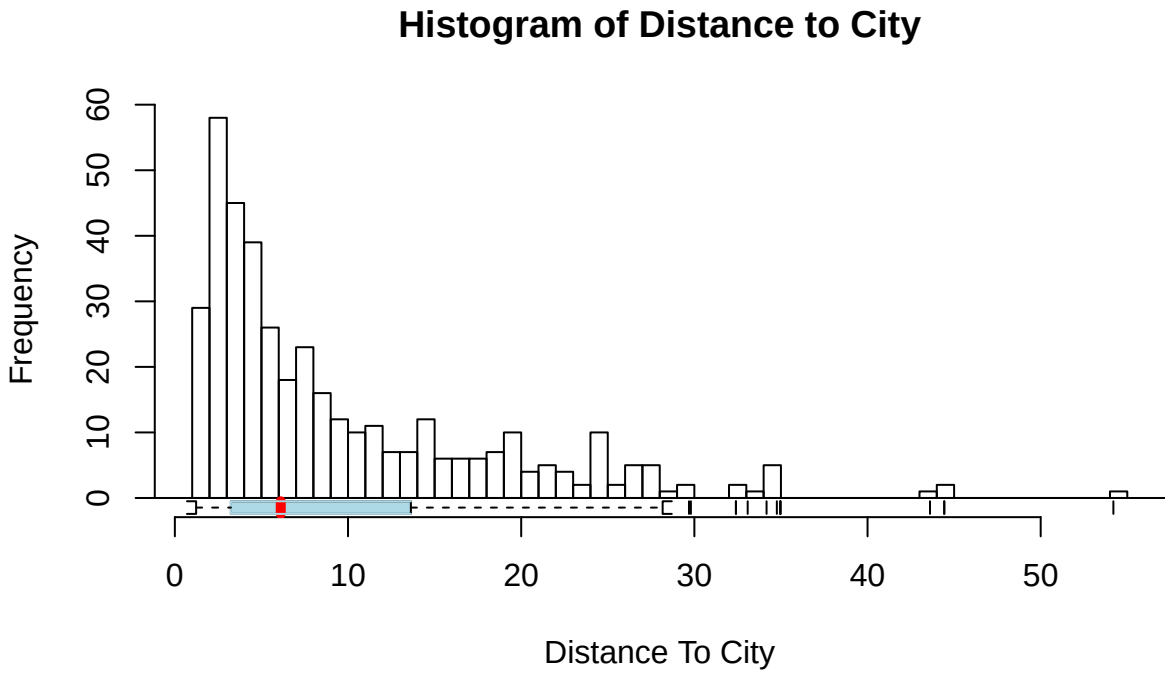
The distribution of *crimeRate_pc* is highly right-skewed with a very long tail. This makes sense as most neighborhoods are very low crime neighborhoods. The distribution of *log(crimeRate_pc)* appears almost bi-modal; however the analysis of skew and kurtosis show a significant improvement of each with a log transformation.

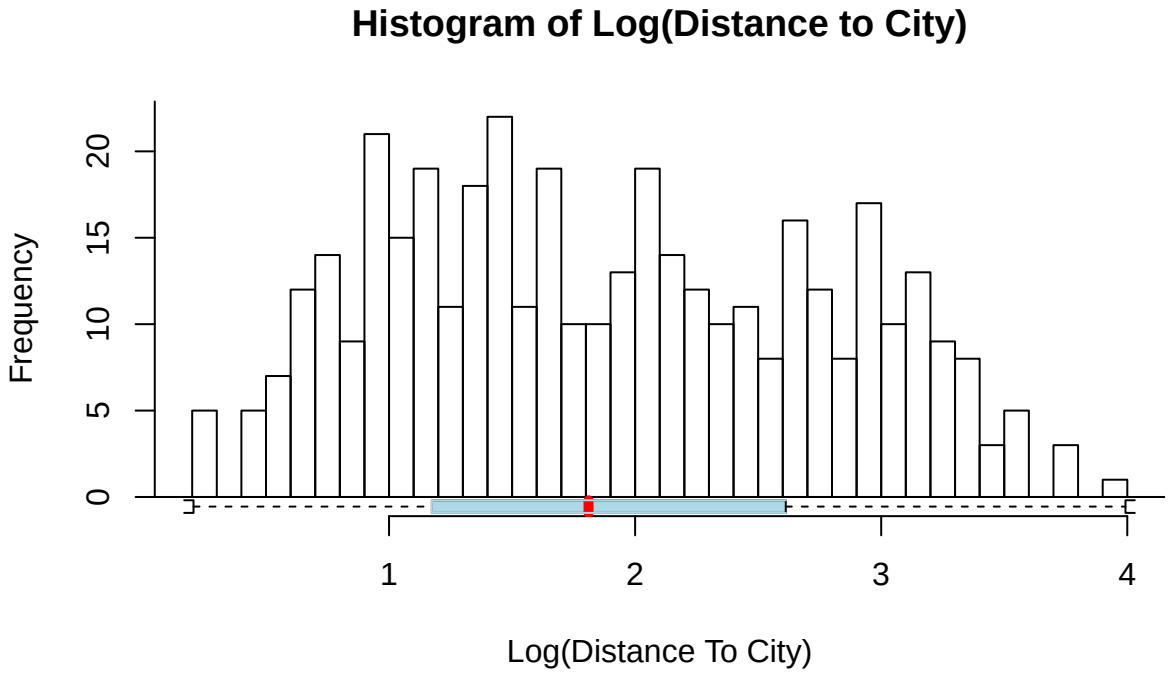




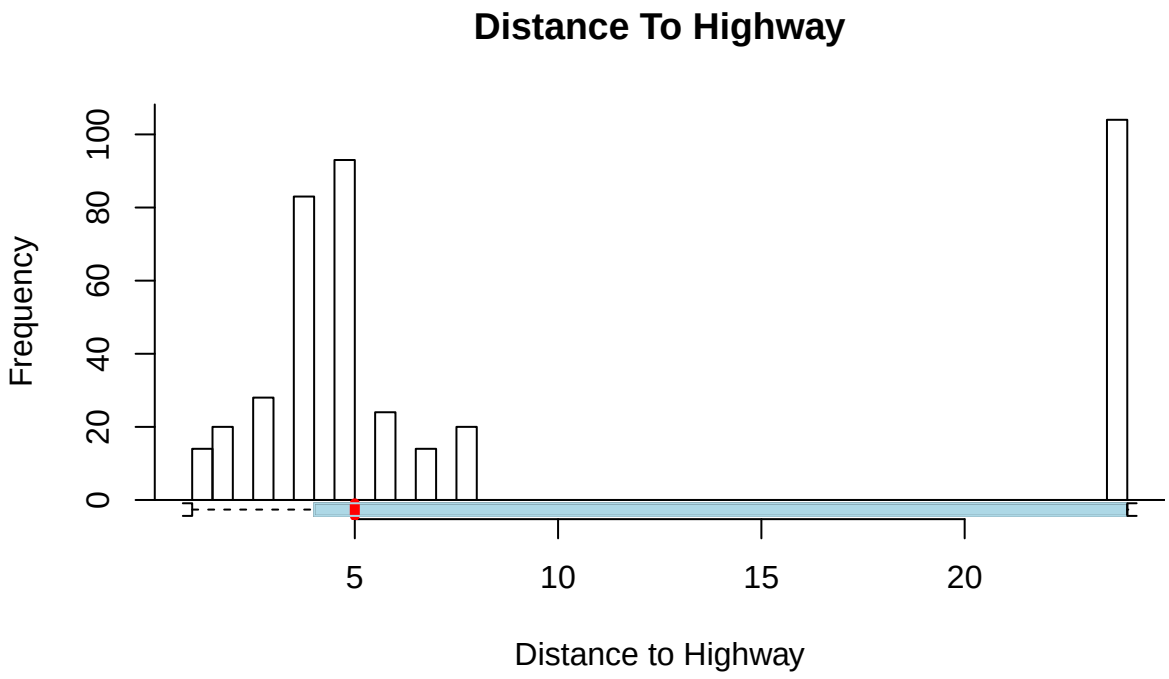
The variable *nonRetailBusiness* is a measure of the footprint of industy in a neaghborhood. This may range from light industrial to manufacturing but that information is not given. The distribution of *nonRetailBusiness* shows a spike at 0.18 business-acres but is otherwise somewhat uniform. There was no transform that improved skew or kurtosis for this variable.

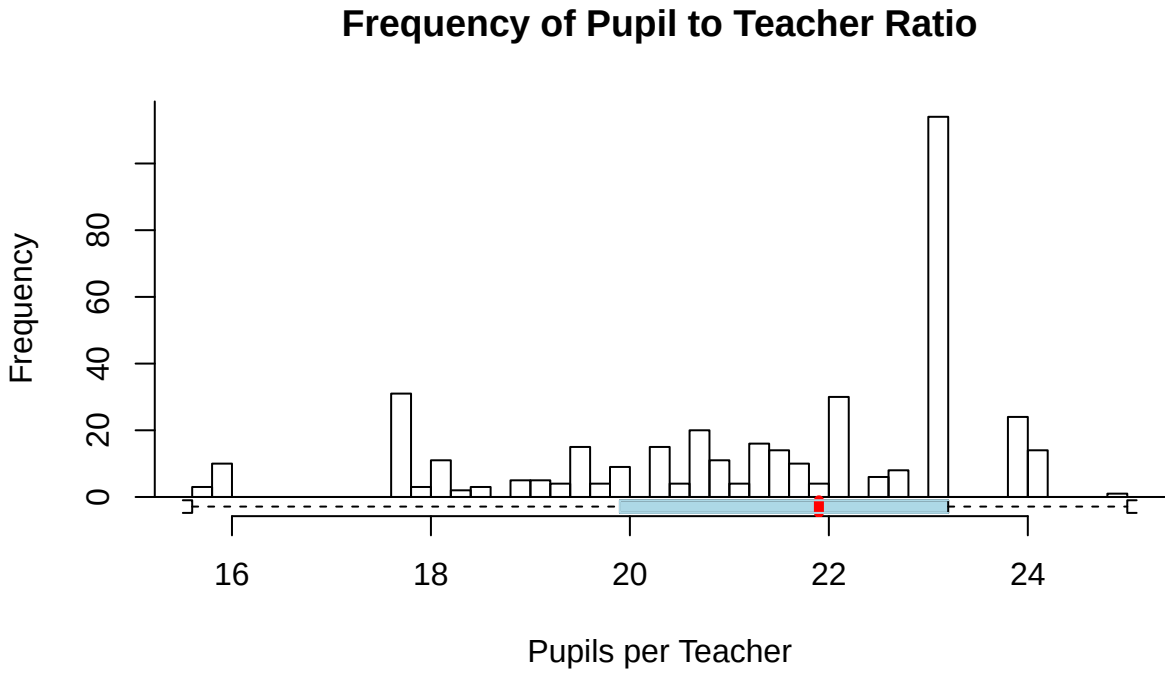
The variable *ageHouse* is the percentage of houses in a neighborhood built before 1950 and shows a significant left-skew with a long tail to the left. However no transformation was found that normalized the skew and kurtosis of this variable.





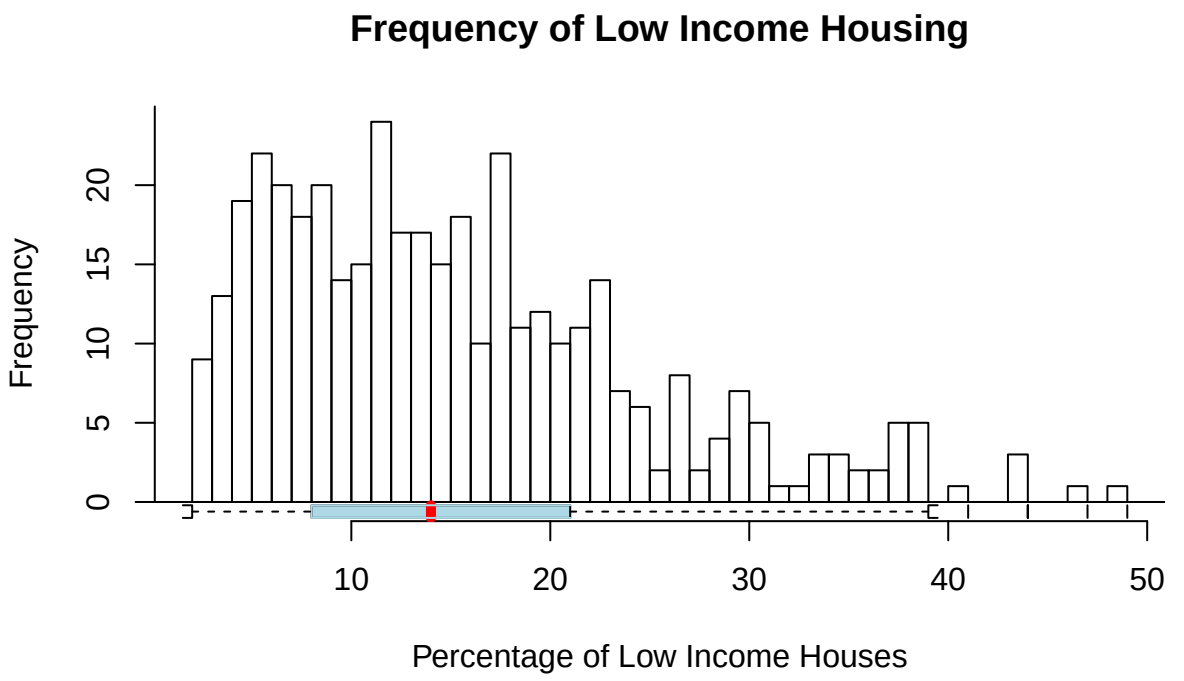
The *distanceToCity* variable shows a right-skewed distrubution that is much improved by a log transformation.

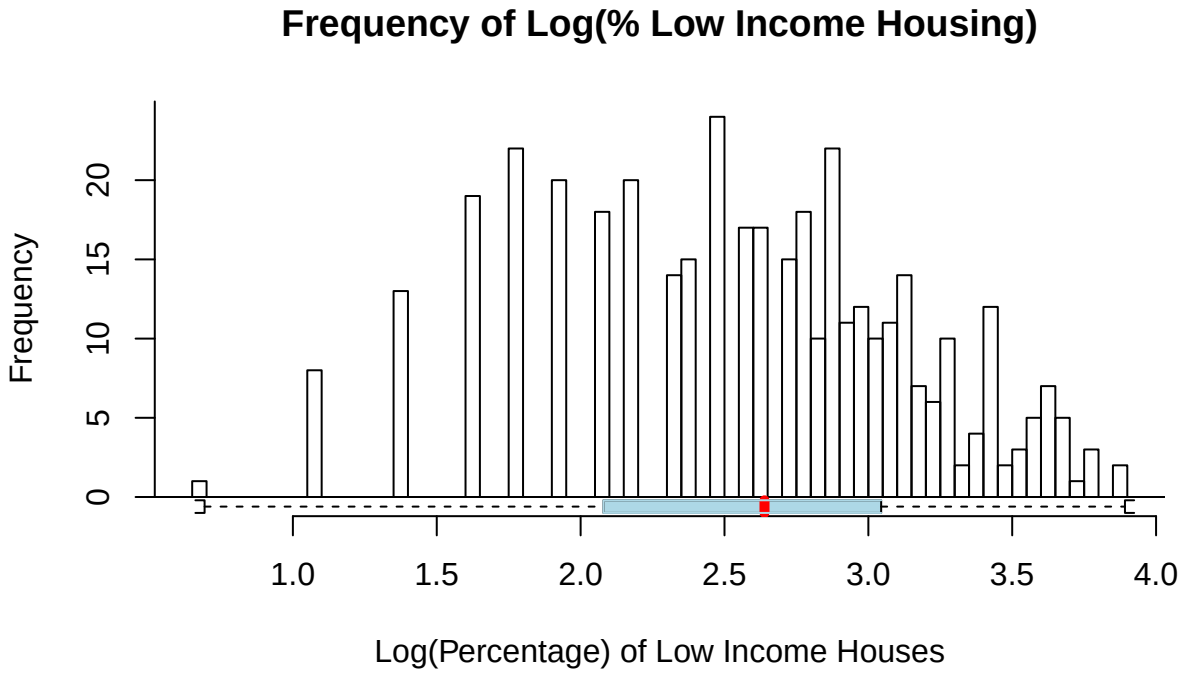




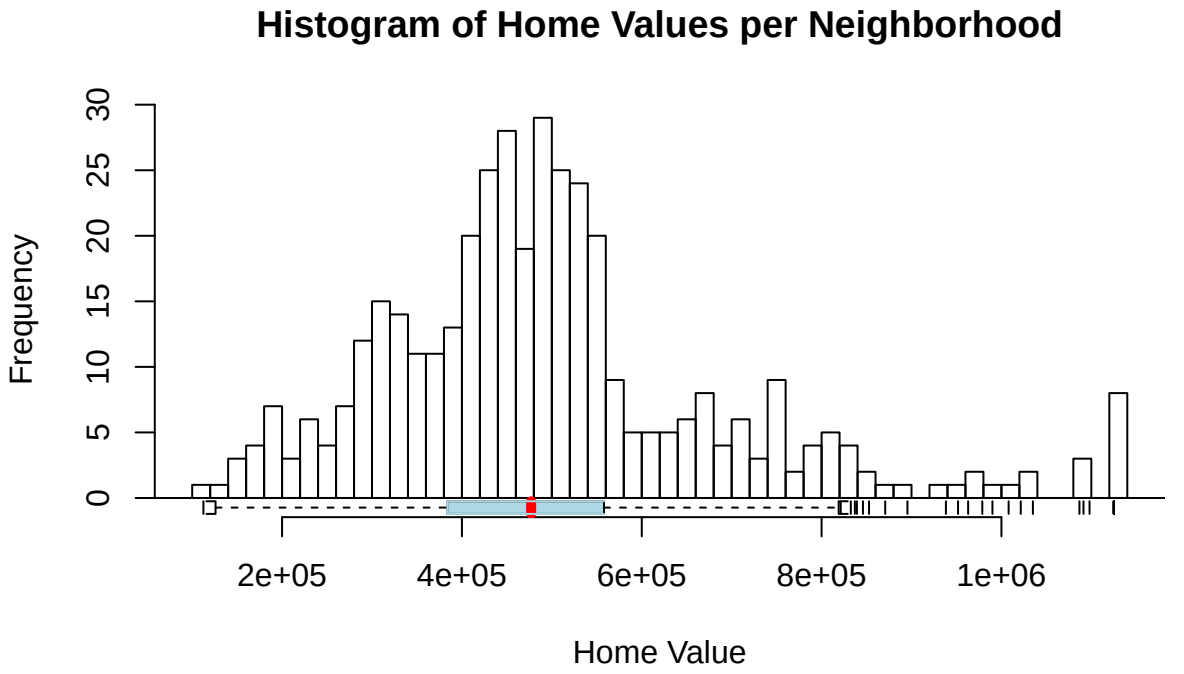
The *distanceToHighway* variable shows the concern with coding error in this histogram as there is a large occurance of the value *24*. About 25% of the dataset have this value, some of which may be correct but it seems unlikely that the *distanceToHighway* variable would be much greater than the *distanceToCity* variable.

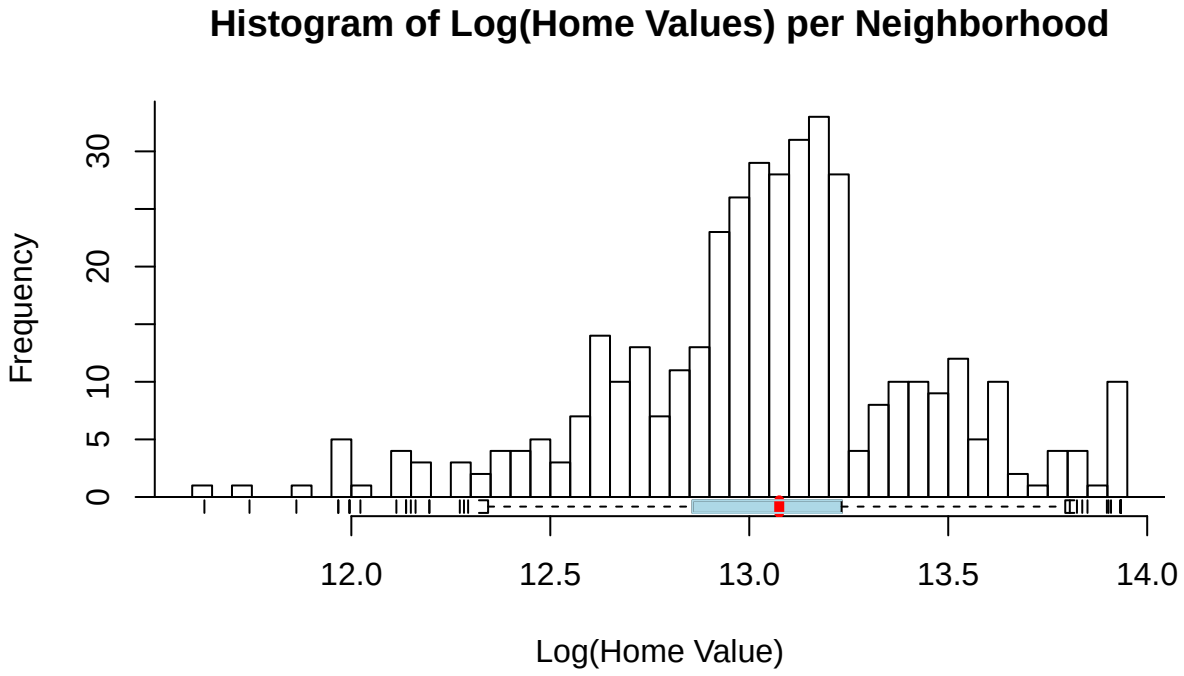
The *pupilTeacherRatio* variable shows a roughly uniform distribution except for a large number of occurances of the value *23*, which must be a more common classroom size.



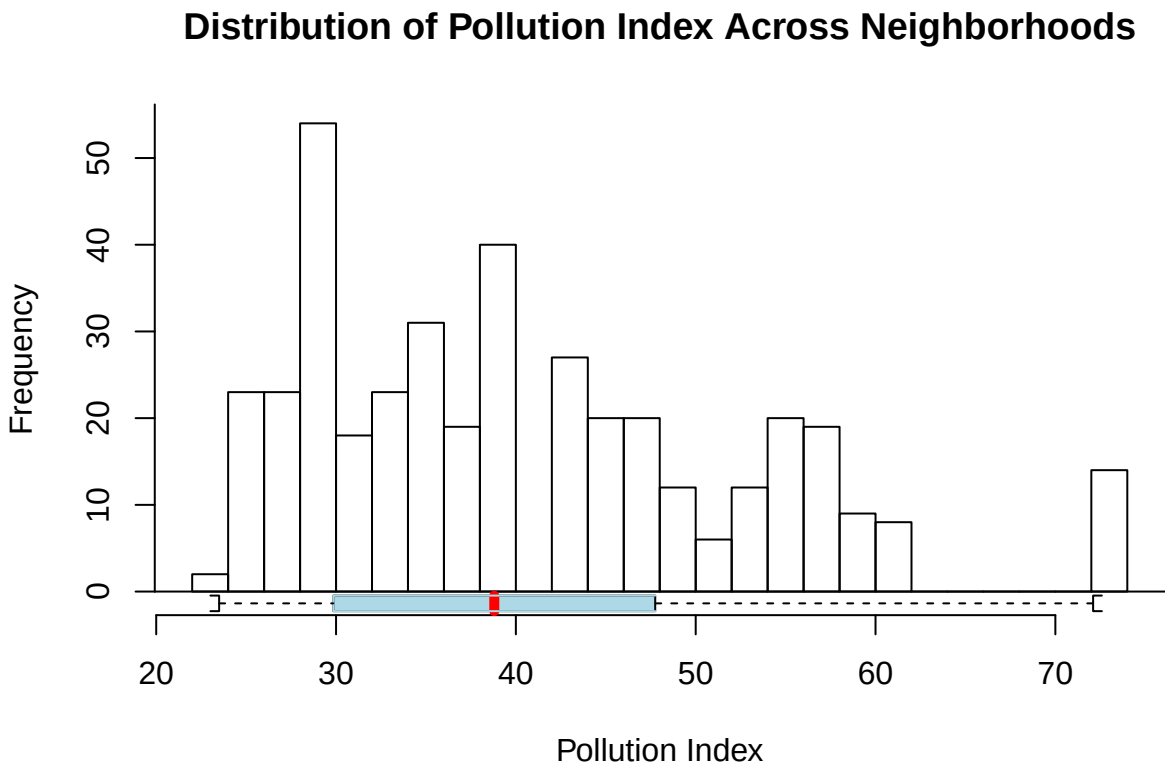


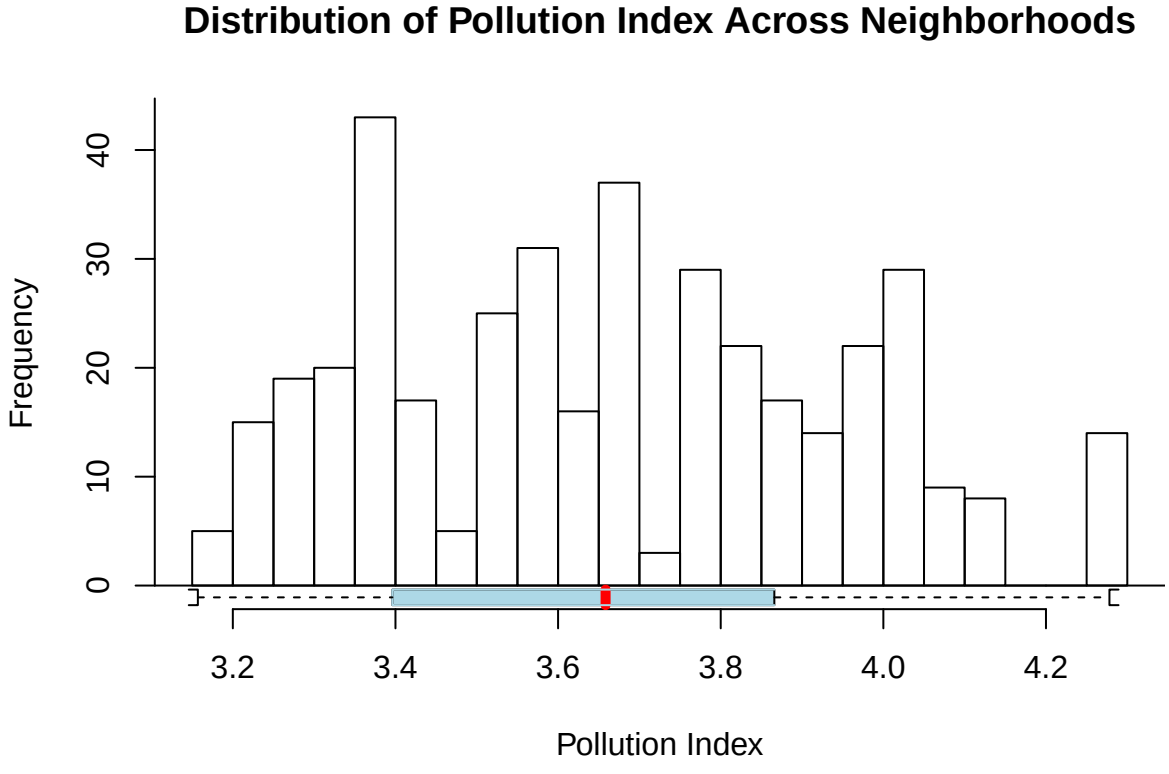
The *pctLowIncome* variable has a right-skewed distribution that tapers off to the right relatively quickly. A log transformation greatly improves the skew and kurtosis of the distribution.



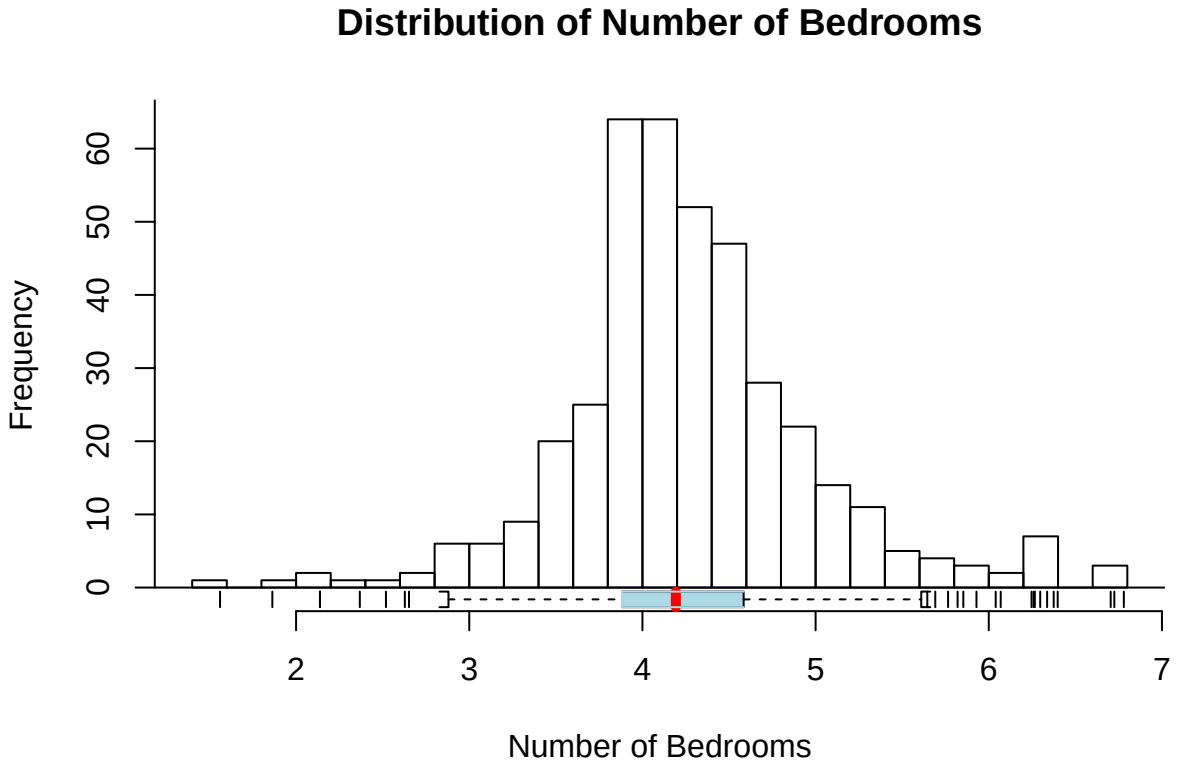


The *homeValue* variable shows a slight right-skew and a log transformation is used to help with this. It also allows discussion in terms of percentage change of home value when controlling for other variables.





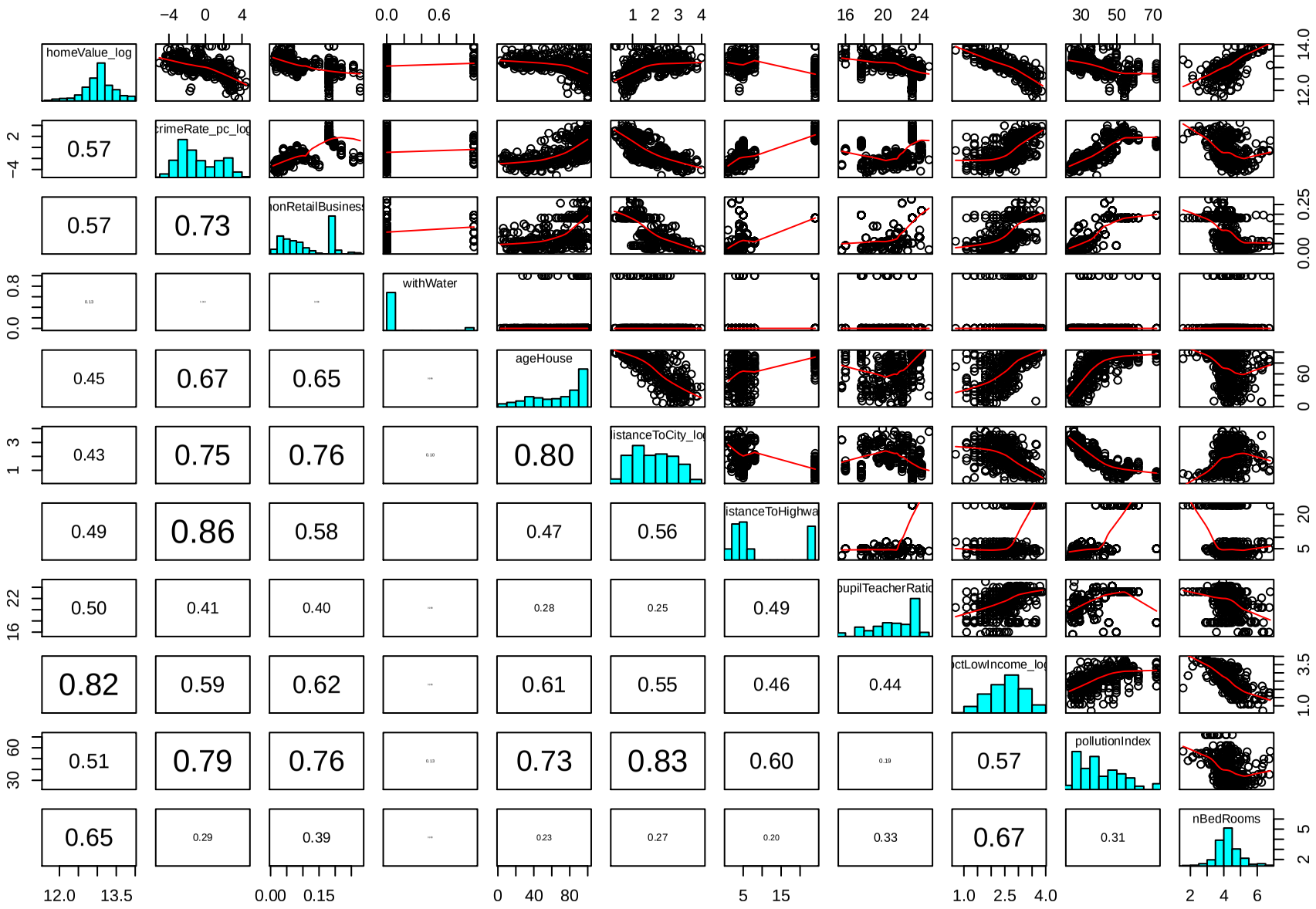
The *pollutionIndex* variable shows a right-skewed distribution upon which we perform a log transform.



The *nBedRooms* variable appears amazingly normal-like in its distribution.

The next page brings all the variables into a single matrix for comparison and to get a first look at correlations to explore further.

Data Set Variable Scatterplot Matrix



DistanceToHighway Variable Detailed Examination

We saw previously that the *distanceToHighway* variable looked suspicious so in this section we look at how to address a possible coding issue. The number of rows in the dataset that have the *distanceToHighway* variable as 24 is 25% of the dataset. Removing these rows would remove a significant amount of data, reducing *N=400* to *N=296*. We examine two strategies and compare them to the row-removal option: replacing values of 24 with the mean of the filtered values or with the value of *distanceToCity*.

The following two tables compare the summaries of the filtered dataset with the summaries of the dataset with transformed values. Comparing the *distanceToHighway_meanMod* and *distanceToHighway_cityMod* shows that the latter is much closer to the values of the filtered dataset. This indicates that replacing the value of 24 with the value of *distanceToCity* is a reasonable transformation to deal with the coding issue. The idea is further substantiated by the proposition that the distance to a city is usually not greater than the distance to a highway as cities are generally located on highways.

The following page shows a set of comparison histograms for the *distanceToHighway* variable with the different transformations. A histogram of *distanceToCity* is included as a reference.

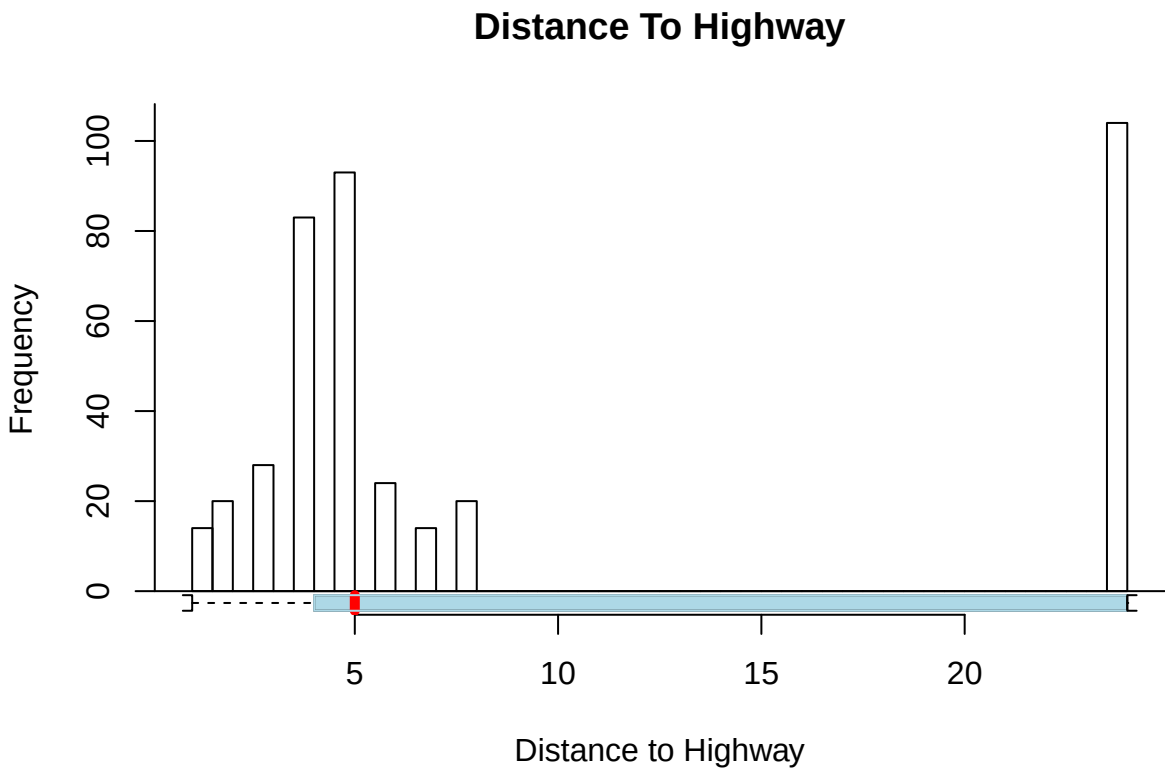
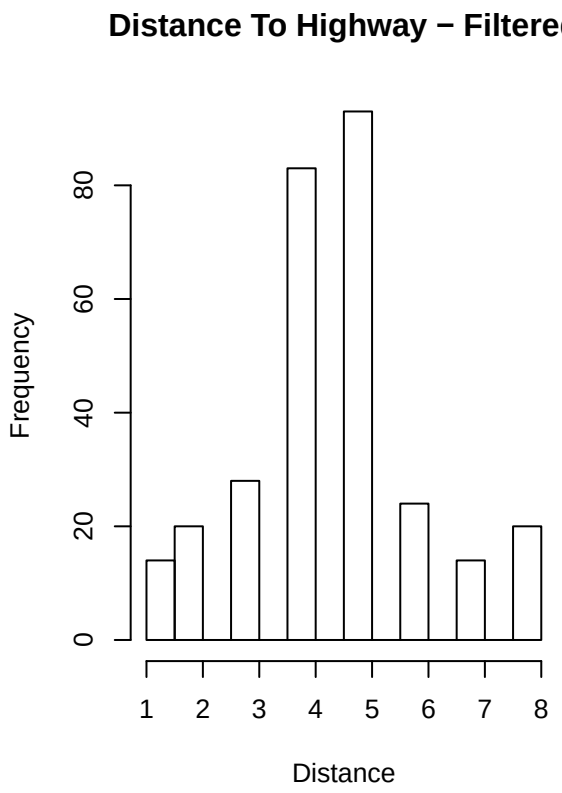


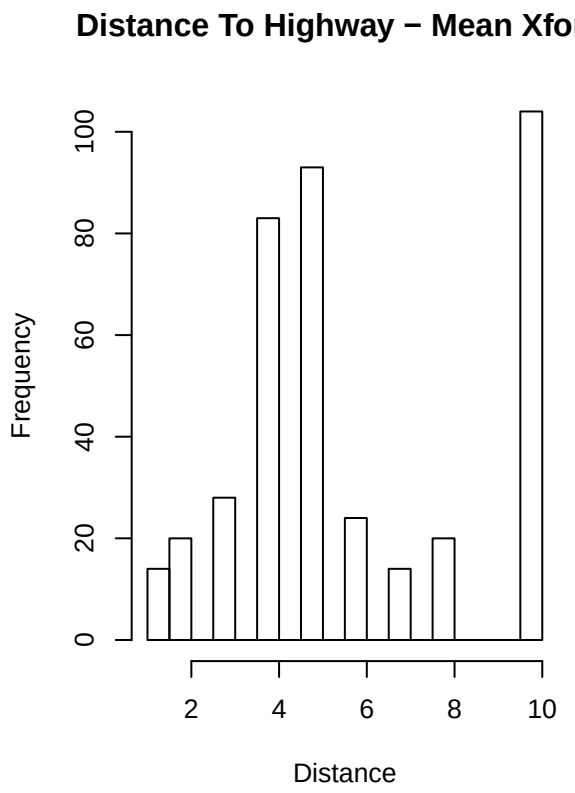
Table 2: Filtered Dataset

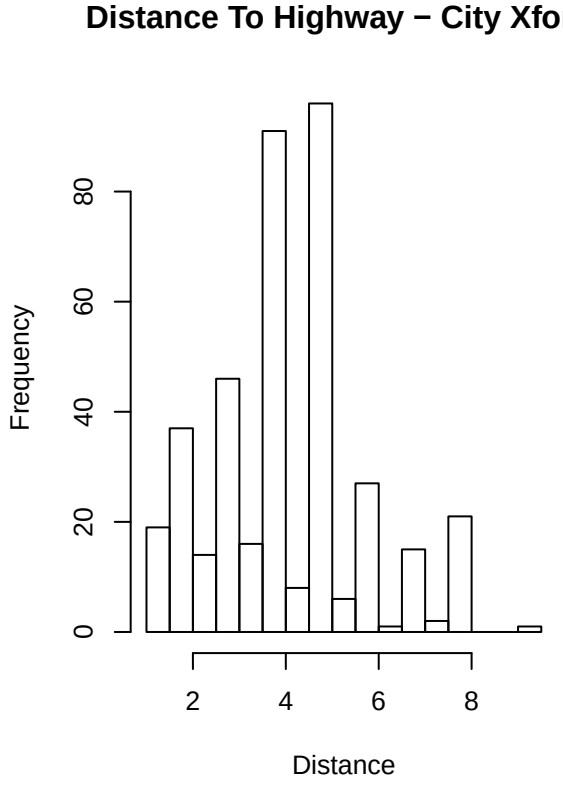
Statistic	N	Mean	St. Dev.	Min	Max
crimeRate_pc	296	0.381	0.610	0.006	3.535
nonRetailBusiness	296	0.087	0.065	0.007	0.277
withWater	296	0.068	0.251	0	1
ageHouse	296	61.367	28.205	2.900	100.000
distanceToCity	296	11.877	9.178	1.562	54.197
distanceToHighway	296	4.517	1.636	1	8
pupilTeacherRatio	296	20.756	2.191	15.600	25.000
pctLowIncome	296	13.037	7.690	2	44
homeValue	296	547,487.300	178,890.300	157,500	1,125,000
pollutionIndex	296	36.438	10.450	23.500	72.100
nBedRooms	296	4.356	0.676	2.903	6.725
crimeRate_pc_log	296	−1.840	1.297	−5.064	1.263
distanceToCity_log	296	2.175	0.802	0.446	3.993
pctLowIncome_log	296	2.403	0.587	0.693	3.784
ageHouse_log	296	3.959	0.643	1.065	4.605
homeValue_log	296	13.165	0.306	11.967	13.933
pollutionIndex_log	296	3.562	0.250	3.157	4.278

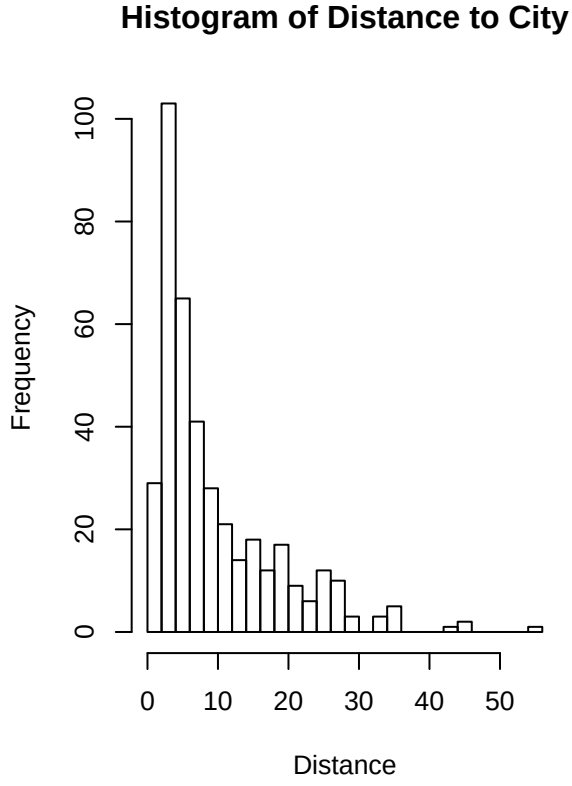
Table 3: Full Dataset

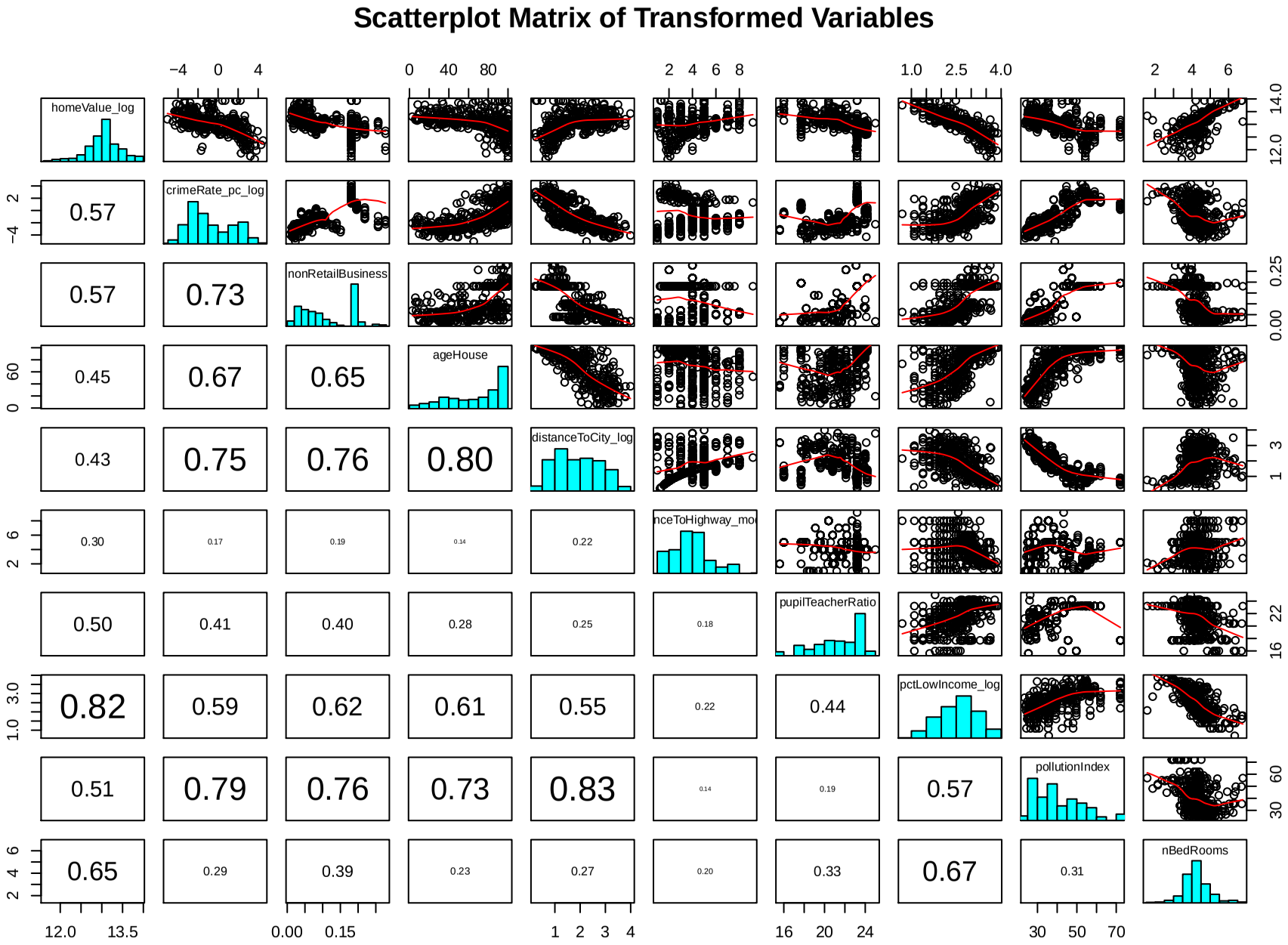
Statistic	N	Mean	St. Dev.	Min	Max
crimeRate_pc	400	3.763	8.872	0.006	88.976
nonRetailBusiness	400	0.112	0.070	0.007	0.277
withWater	400	0.068	0.251	0	1
ageHouse	400	68.932	27.977	2.900	100.000
distanceToCity	400	9.638	8.786	1.228	54.197
distanceToHighway	400	9.582	8.672	1	24
pupilTeacherRatio	400	21.391	2.168	15.600	25.000
pctLowIncome	400	15.795	9.341	2	49
homeValue	400	499,584.400	196,115.700	112,500	1,125,000
pollutionIndex	400	40.615	11.825	23.500	72.100
nBedRooms	400	4.266	0.719	1.561	6.780
crimeRate_pc_log	400	−0.763	2.164	−5.064	4.488
distanceToCity_log	400	1.892	0.868	0.205	3.993
pctLowIncome_log	400	2.577	0.631	0.693	3.892
ageHouse_log	400	4.099	0.605	1.065	4.605
homeValue_log	400	13.046	0.397	11.631	13.933
pollutionIndex_log	400	3.664	0.282	3.157	4.278
distanceToHighway_modMean	400	5.834	2.632	1.000	9.582
distanceToHighway_modCity	400	4.192	1.698	1.000	9.159


Distance To Highway − Filtered
Frequency
0
20
40
60
80
1
2
3
4
5
6
7
8
Distance


Distance To Highway − Mean Xform
Frequency
0
20
40
60
80
100
2
4
6
8
10
Distance

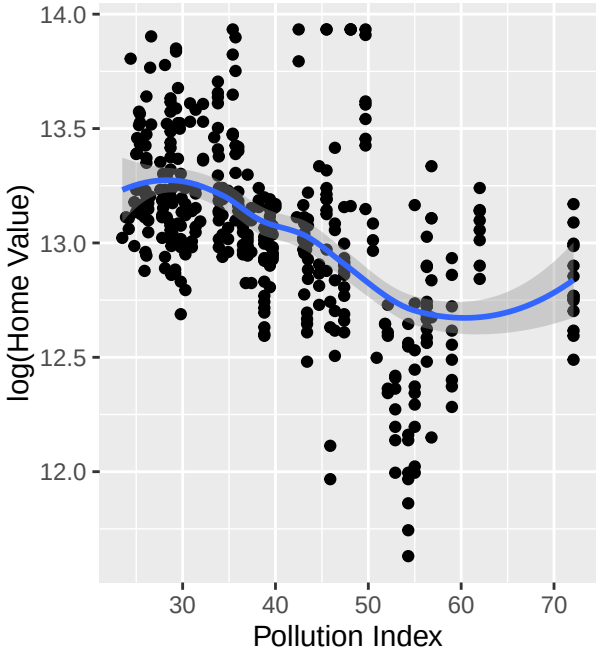

Distance To Highway − City Xform
Frequency
0
20
40
60
80
2
4
6
8
Distance

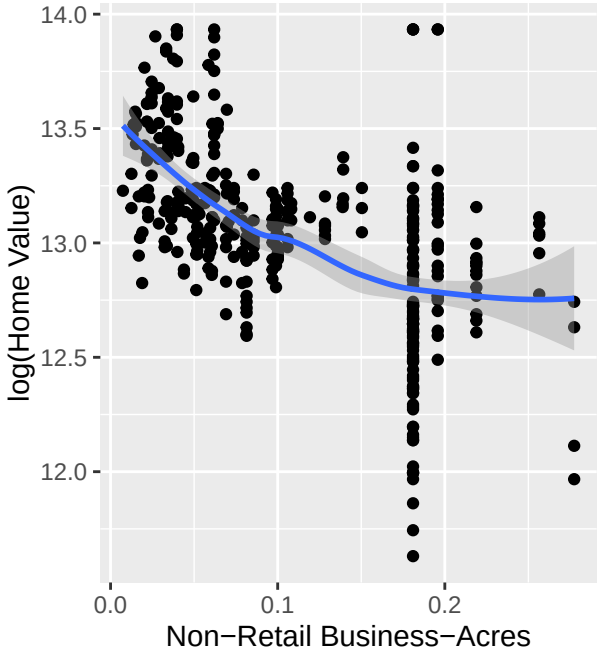

Histogram of Distance to City
Frequency
0
20
40
60
80
100
0
10
20
30
40
50
Distance

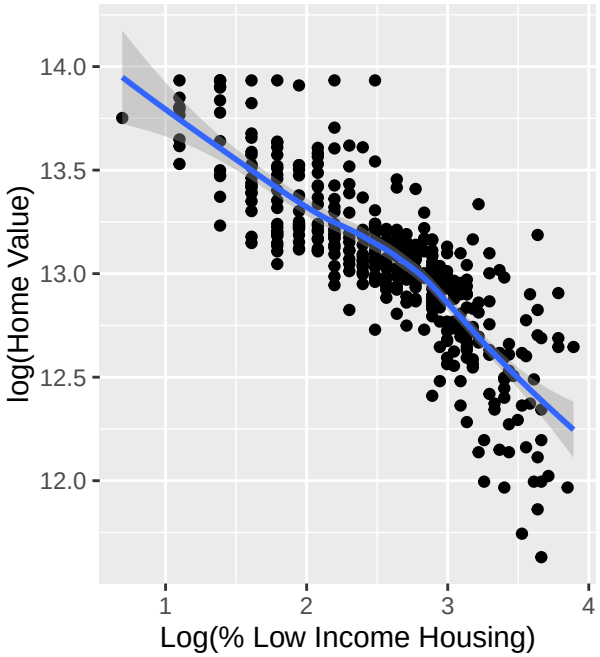

Scatterplot Matrix of Transformed Variables
homeValue_log
crimeRate_pc_log
nonRetailBusiness
ageHouse
distanceToCity_log
nceToHighway_mo
pupilTeacherRatio
pctLowIncome_log
pollutionIndex
nBedRooms
0.57
0.57
0.73
0.45
0.67
0.65
0.43
0.75
0.76
0.80
0.30
0.17
0.19
0.14
0.22
0.50
0.41
0.40
0.28
0.25
0.18
0.82
0.59
0.62
0.61
0.55
0.22
0.44
0.51
0.79
0.76
0.73
0.83
0.14
0.19
0.57
0.65
0.29
0.39
0.23
0.27
0.20
0.33
0.67
0.31

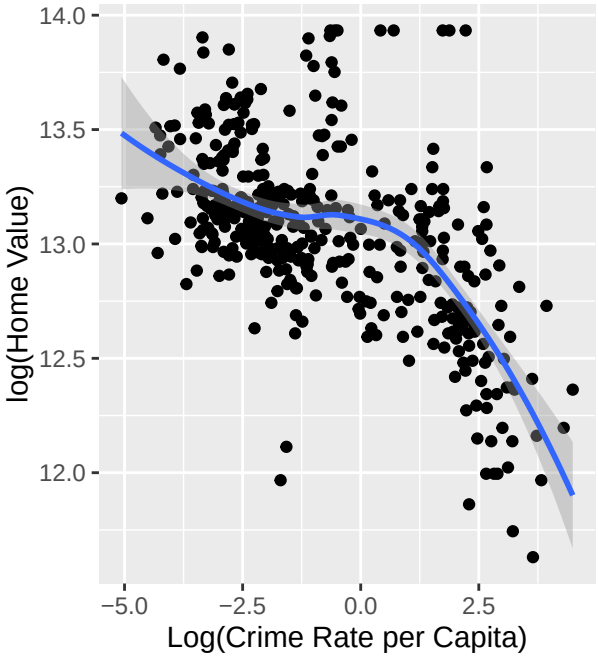
Multivariate Data Analysis

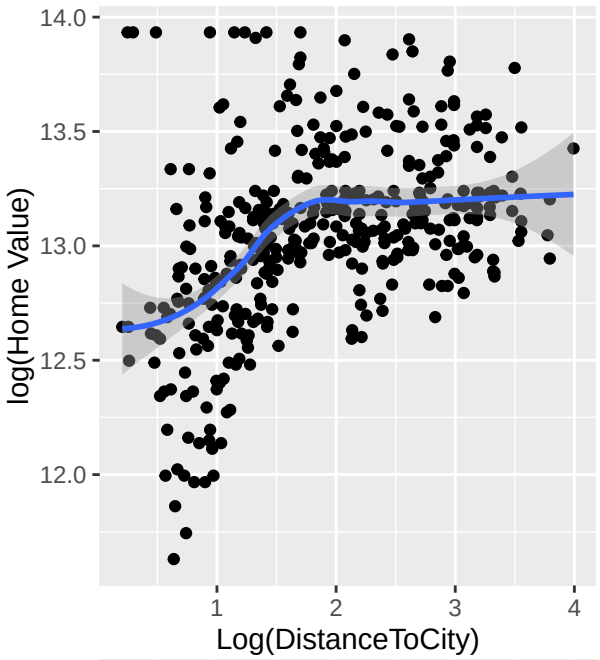
First we will examine the relationships of the environment variables on home values graphically. We can see that there are definitely relationships between most of the environmental variables and home values, as shown on the next two pages of graphs.

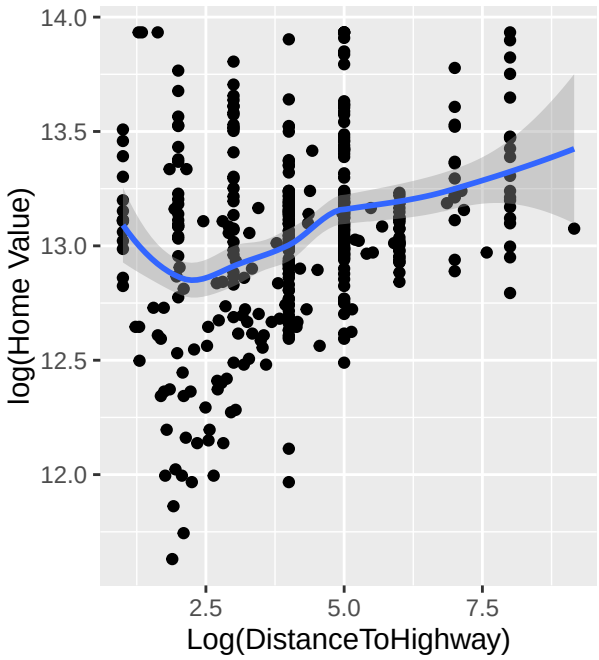


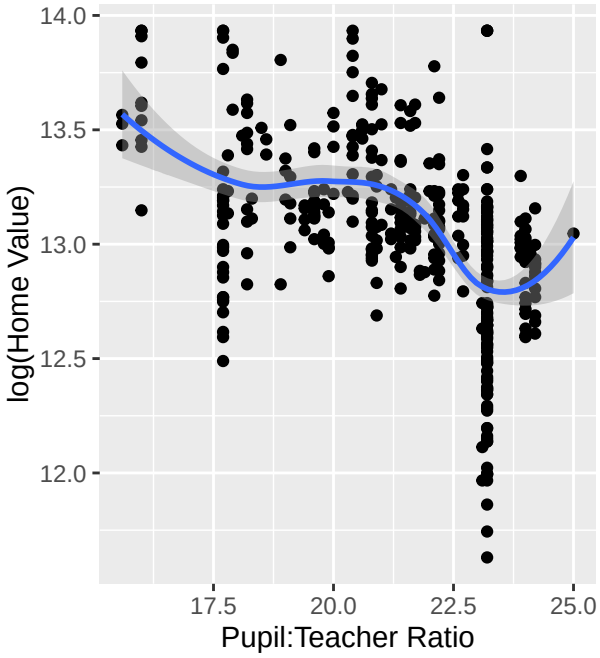




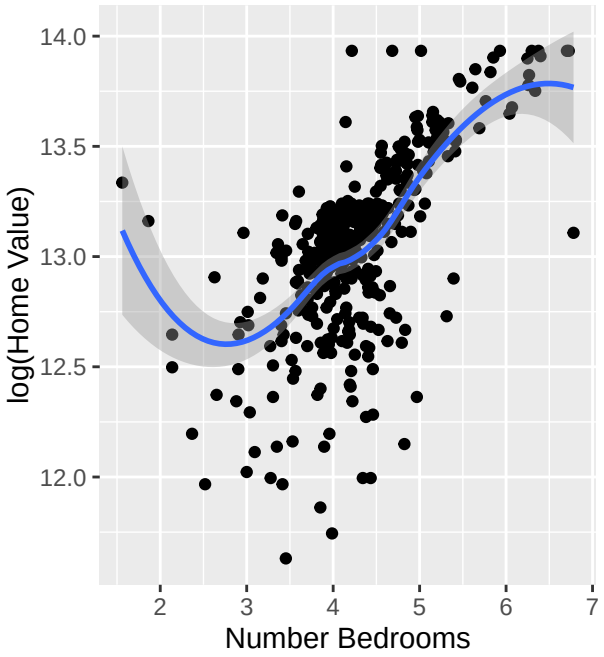


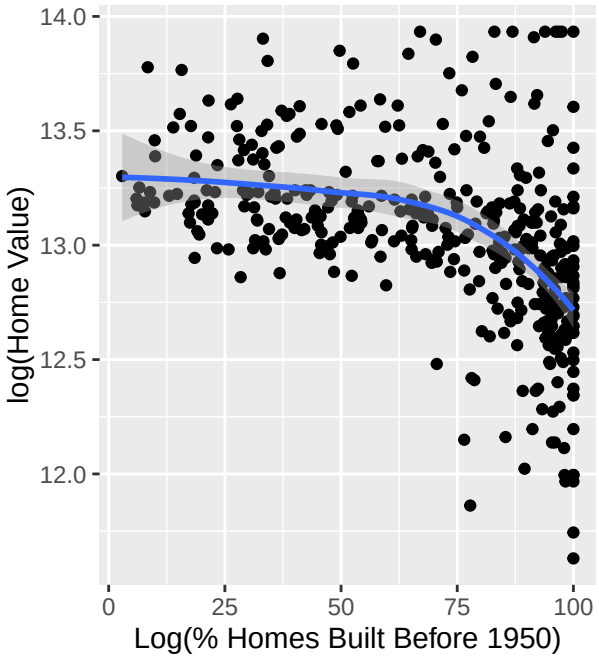

14.0
13.5
13.0
12.5
12.0
log(Home Value)
1
2
3
4
Log(DistanceToCity)


14.0
13.5
13.0
12.5
12.0
log(Home Value)
2.5
5.0
7.5
Log(DistanceToHighway)


14.0
13.5
13.0
12.5
12.0
log(Home Value)
17.5
20.0
22.5
25.0
Pupil:Teacher Ratio

These final two graphs show relationships between home attributes and home values.





Models Incorporating Environment Variables

Put something here about all the different models that were tried.

Include models with interaction terms

Discuss omitted variables and how they bias the models

What hypotheses can we test?

Choose the best model(s) and explain the parameters for them.

Table 4: Regression Model Comparison

	Dependent variable:				
	log(Home Values)				
	(1)	(2)	(3)	(4)	(5)
log(Low Income Housing)	-0.518^{***}	-0.469^{***}	-0.474^{***}	-0.461^{***}	-0.448^{***}
	(0.018)	(0.022)	(0.022)	(0.022)	(0.022)
log(Crime Rate)		-0.024^{***}	-0.030^{***}	-0.026^{***}	-0.024^{***}
		(0.006)	(0.009)	(0.006)	(0.006)
Pollution Index			0.002		
			(0.002)		
Close To Water					0.027^{***}
					(0.007)
Distance To Highway				0.144^{***}	0.137^{***}
				(0.044)	(0.043)
Constant	14.380^{***}	14.238^{***}	14.181^{***}	14.205^{***}	14.060^{***}
	(0.048)	(0.061)	(0.082)	(0.061)	(0.069)
Observations	400	400	400	400	400
R^2	0.677	0.688	0.689	0.696	0.709
Adjusted R^2	0.676	0.686	0.686	0.694	0.706
Residual Std. Error	0.226	0.222	0.222	0.220	0.215
F Statistic	834.337^{***}	437.399^{***}	291.964^{***}	302.291^{***}	240.388^{***}

Note: *p<0.1; **p<0.05; ***p<0.01

Part 2 - Modeling and Forecasting a Real-World Macroeconomic Financial Time Series

Exploratory Data Analysis

In order to better understand the time series and analyze the possible underlying processes we must first observe and explore the time series. The first set of plots reveals:

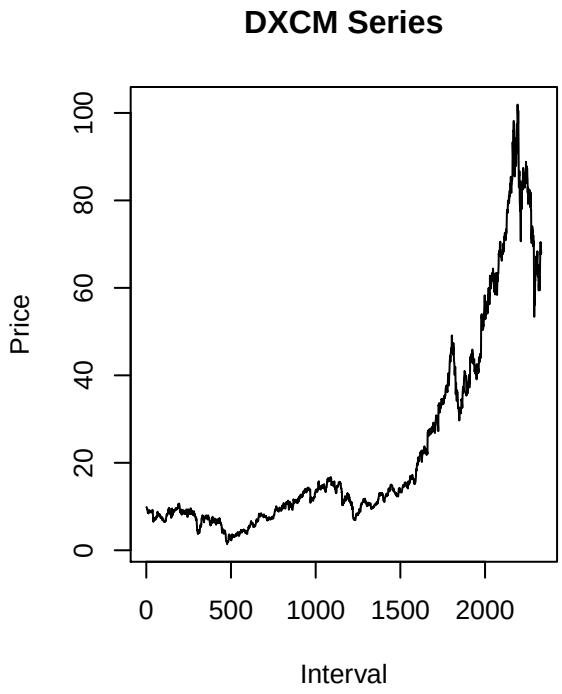
- The series is non-stationary; it has a persistent upward trend;
- There are shocks at approximately time periods 500, 1200, 1800 and 2200;
- The autocorrelation shows a very slight decay over the entire correlagram;
- The partial autocorrelation shows barely significant results at lags 14 and 32;
- There doesn't appear to be any seasonality in the time series;
- We do not know the frequency of the time series;
- The series is of closing prices of DXCM

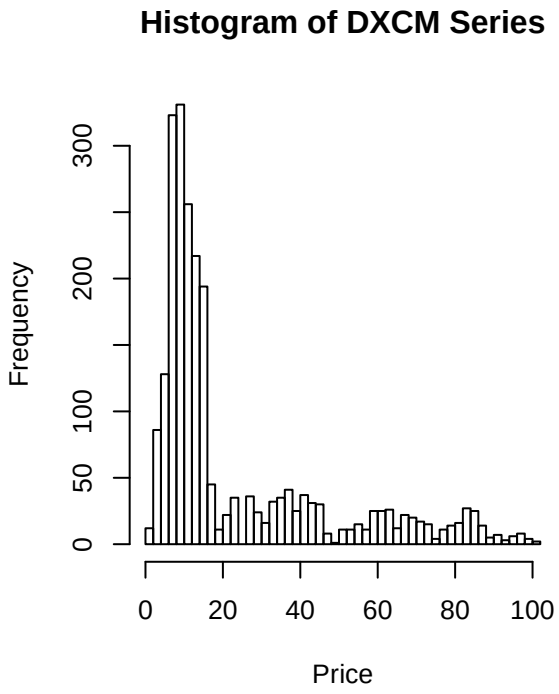
To remove the trend from the series we take the first difference and replot to check the results. In the differenced series we observe:

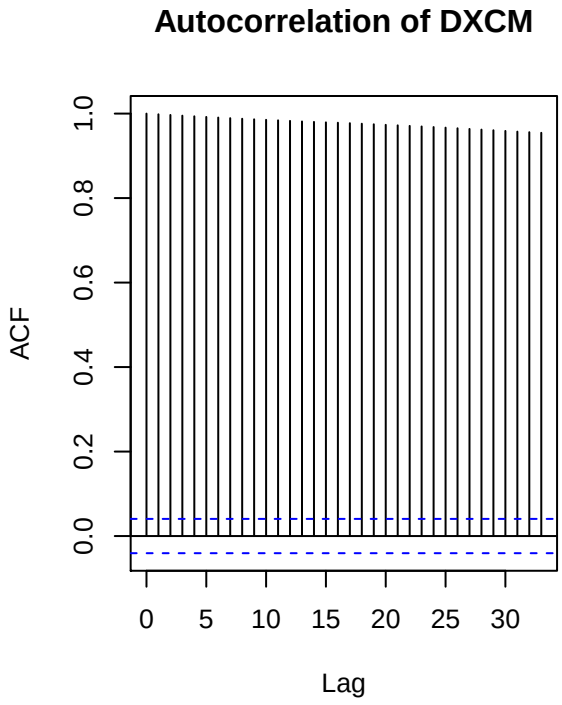
- The first difference series has a more or less white noise appearance until approximately time interval 1600 where the volatility of the series increases dramatically. This corresponds to the sudden, persistent upward trend in the original series.
- The autocorrelation shows marginally significant results at lags 13, 15, 16, 24, 31
- The partial autocorrelation shows a cyclic behavior that doesn't appear to decline, with significant results at lags 11, 13, 14, 15, 16, 24, 25, 31

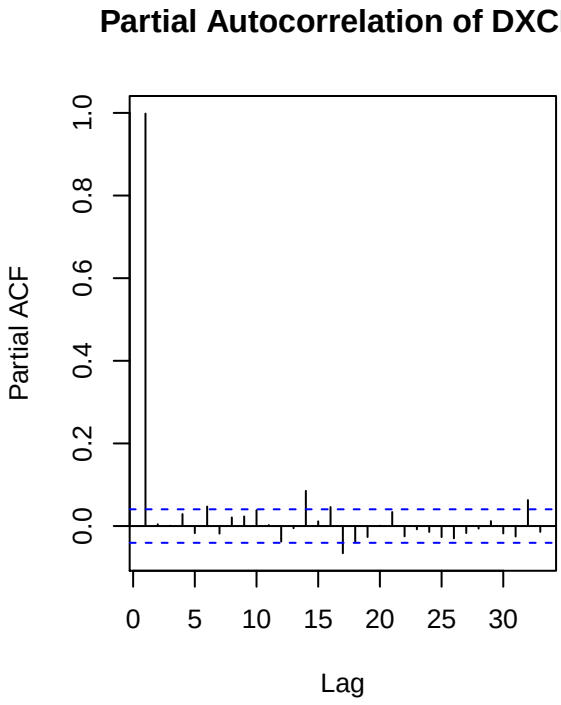
We also examine the first difference of the log of the series and replot to check results. In the differenced log series we observe:

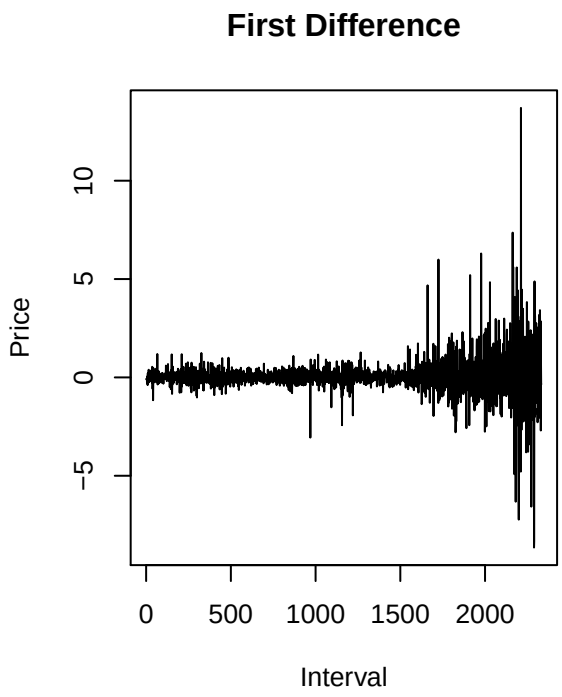
- The volatility appears to be reversed such that it is around interval 300-500, and overall the volatility of the differenced log series is higher.
- The ACF shows only a small results at lag 15 and 20.
- The PACF shows a cyclic behavior with significant results at lags 9, 15, 20

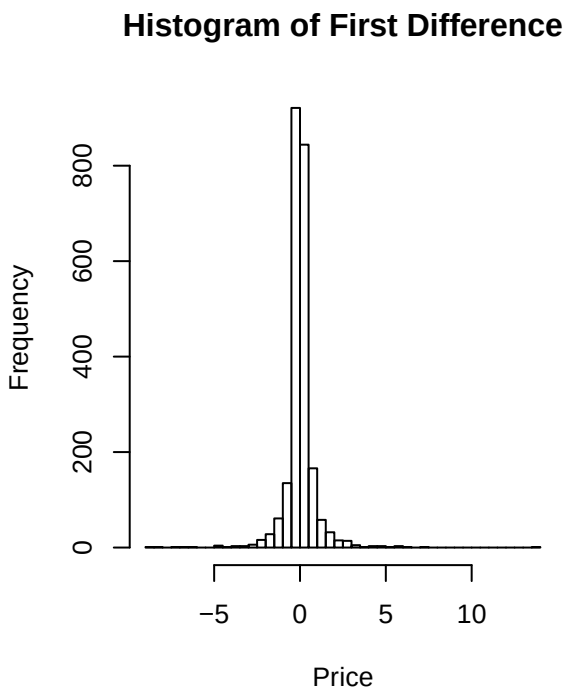

DXCM Series
Price
0
20
40
60
80
100
0
500
1000
1500
2000
Interval

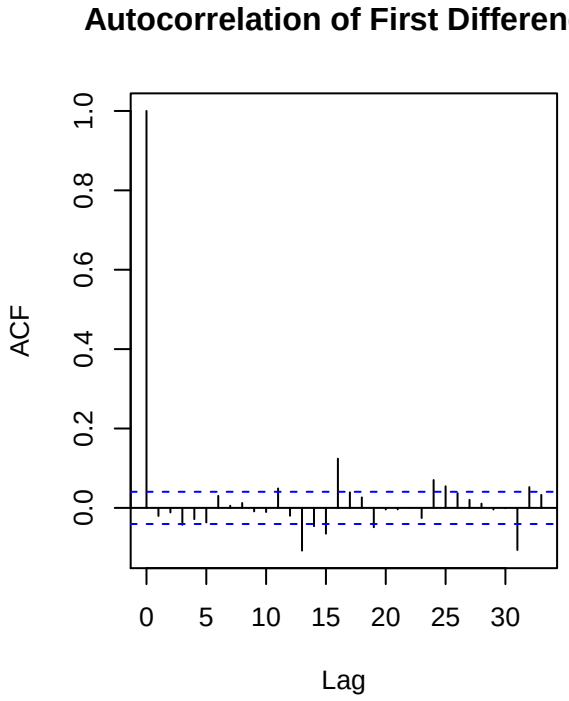

Histogram of DXCM Series
Frequency
0
50
100
200
300
0
20
40
60
80
100
Price

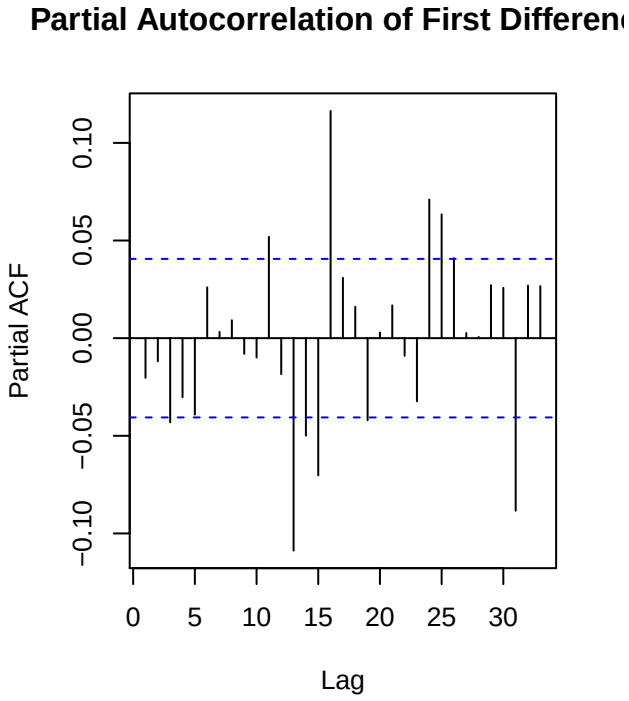

Autocorrelation of DXCM
ACF
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
Lag

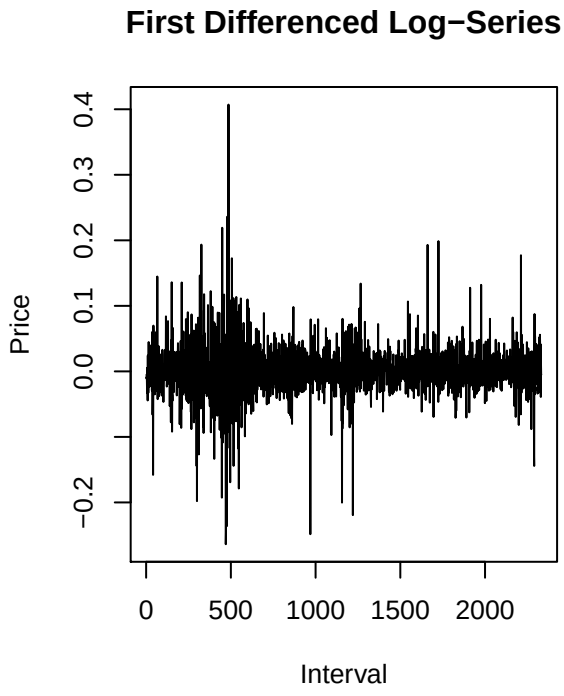

Partial Autocorrelation of DXCM
Partial ACF
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
Lag

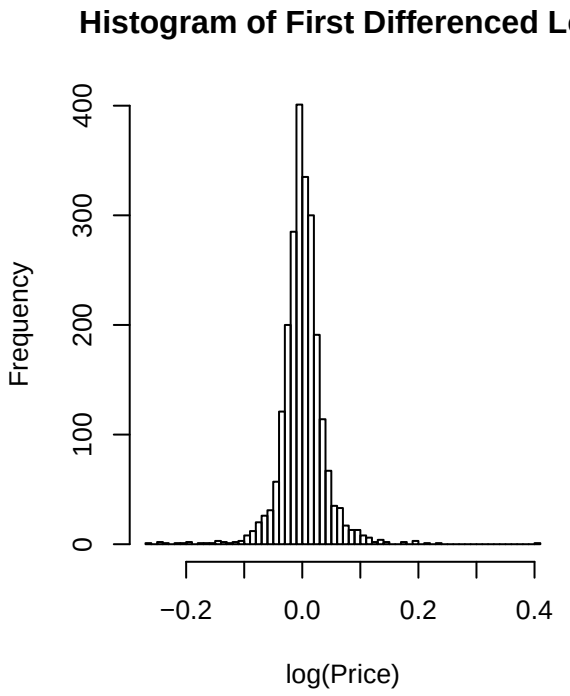

First Difference
Price
Interval
10
5
0
−5
0
500
1000
1500
2000

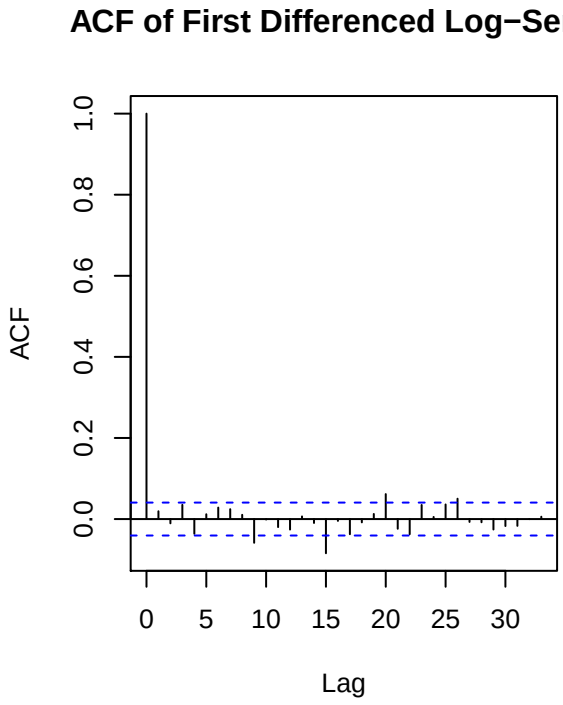

Histogram of First Difference
Frequency
Price
800
600
400
200
0
−5
0
5
10

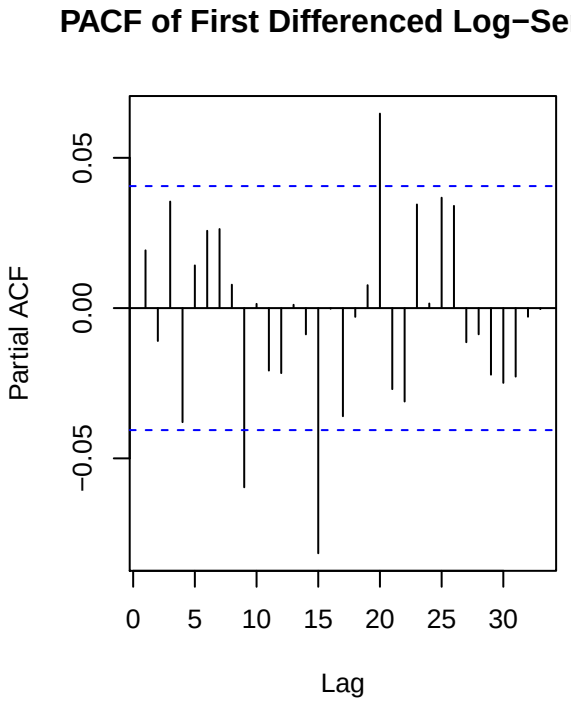

Autocorrelation of First Difference
ACF
Lag
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30


Partial Autocorrelation of First Differen
Partial ACF
Lag
0.10
0.05
0.00
−0.05
−0.10
0
5
10
15
20
25
30


First Differenced Log−Series
Price
−0.2
0.0
0.1
0.2
0.3
0.4
0
500
1000
1500
2000
Interval


Histogram of First Differenced Log
Frequency
0
100
200
300
400
−0.2
0.0
0.2
0.4
log(Price)


ACF of First Differenced Log−Series
ACF
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
Lag


PACF of First Differenced Log−Series
Partial ACF
−0.05
0.00
0.05
0
5
10
15
20
25
30
Lag

Part 3 - Forecast Web Search Activity for "Global Warming"

Part 4 - Forecast Inflation-Adjusted Gas Price