Homework 2

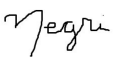
Group 3
Nathalia Negri
Troy Yang

305-318-3446
781-812-4931

negri.n@husky.neu.edu
yang.tr@husky.neu.edu

Percentage of Effort Contributed by Student 1:_____50%_____

Percentage of Effort Contributed by Student 2:_____50%_____

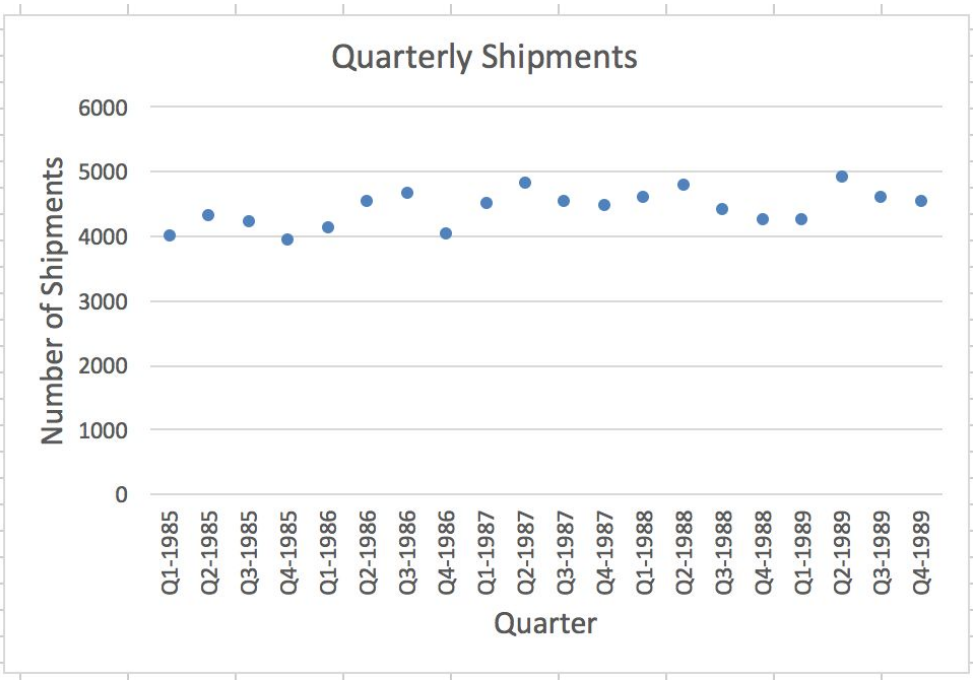
Signature of Student 1:_____  **_____**

Signature of Student 2:_____  **_____**

Submission Date:_____May 29, 2017_____

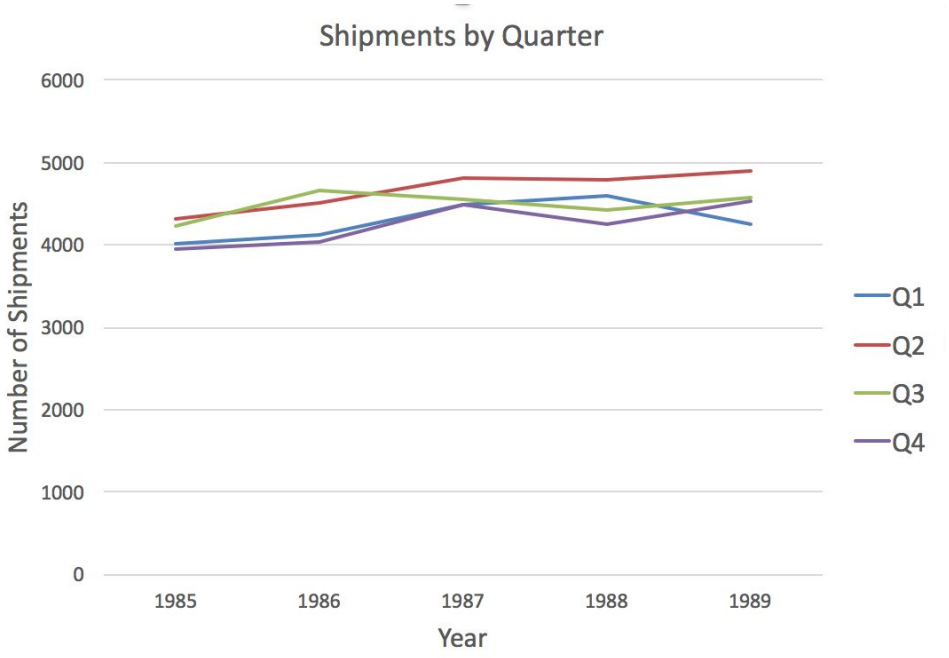
Problem 3.1 15 points (see 3.1_ApplianceShipments.xlsx for data set)

a. Scatter plot of data

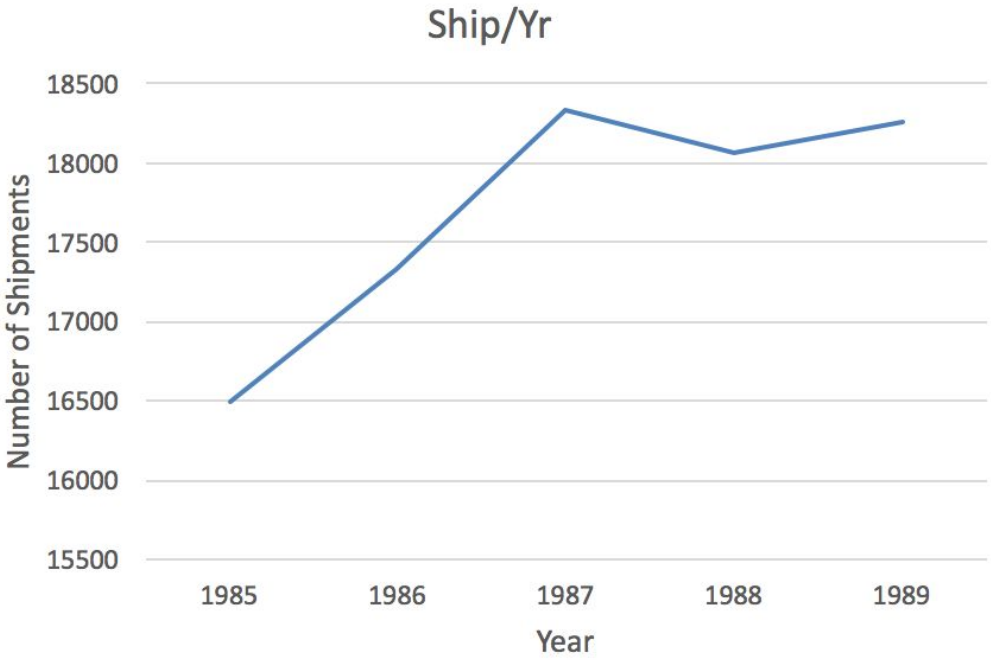


b. Typically Q1 and Q4 are lower than Q2 and Q3. There is obvious seasonality with an increase in shipments during Q2 and Q3.

c. As you can see below, Q2 and Q3 generally have a higher number of shipments, which backs our assumption based off the initial scatter plot.

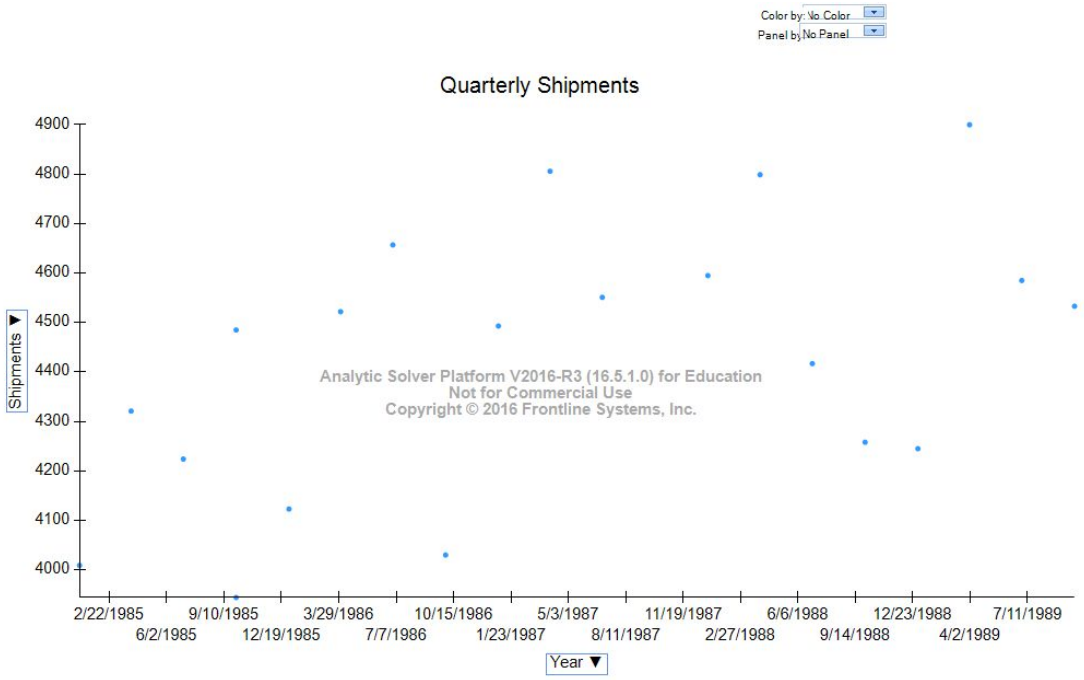


d. Line graph at yearly aggregated level

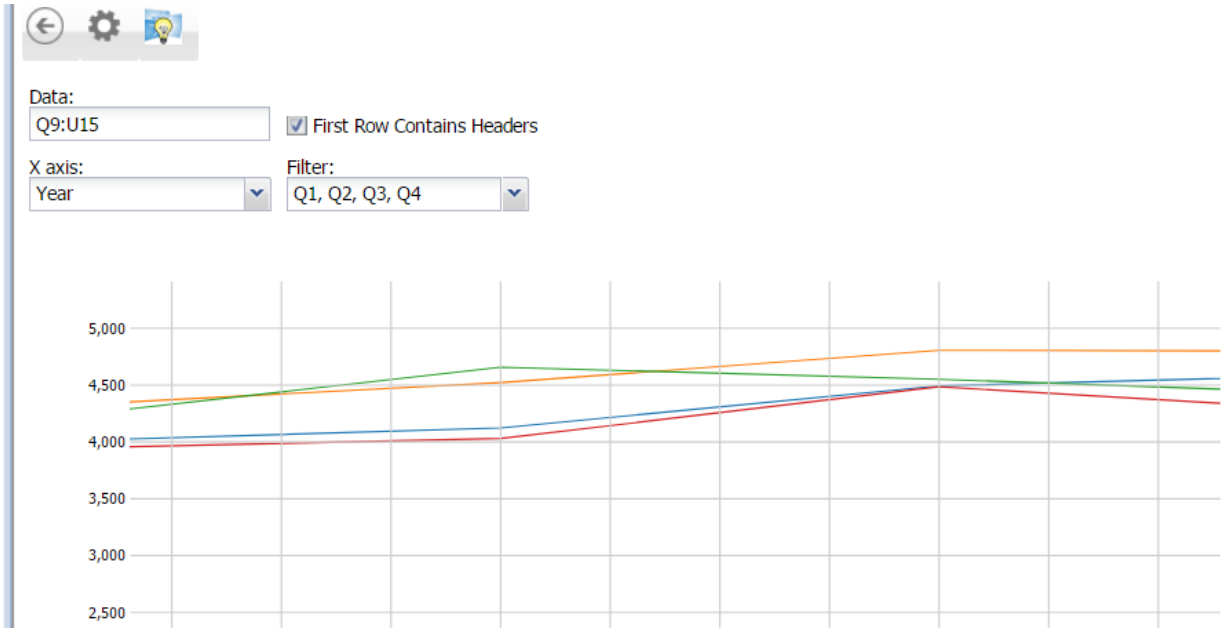


e.

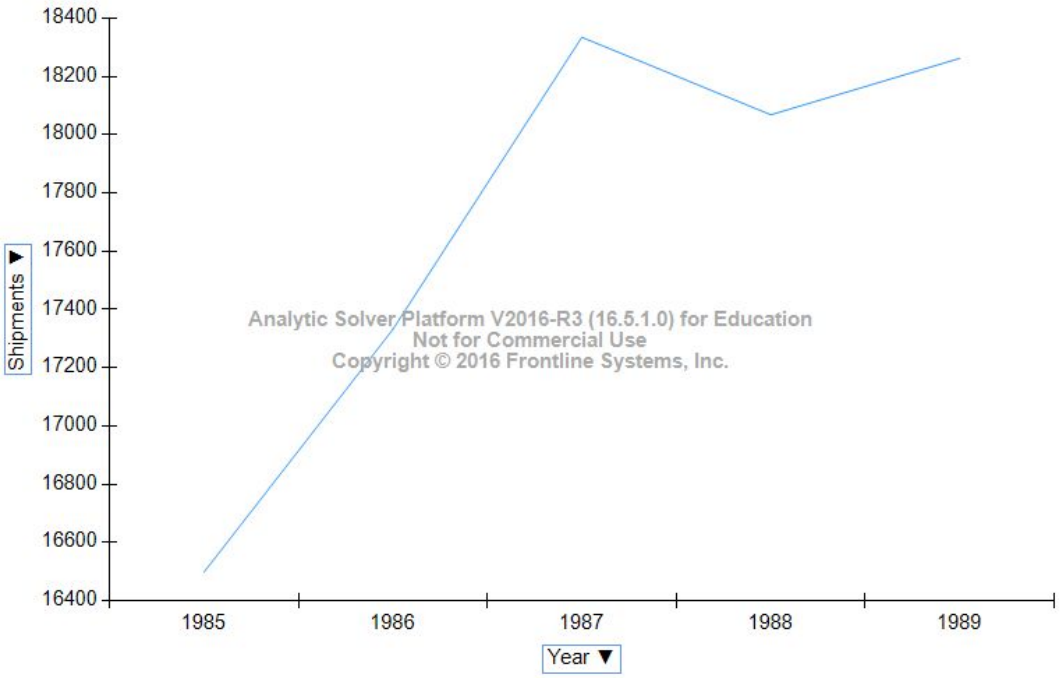
i. First Plot Recreated



ii. Second Plot Recreated



iii. Third Plot Recreated

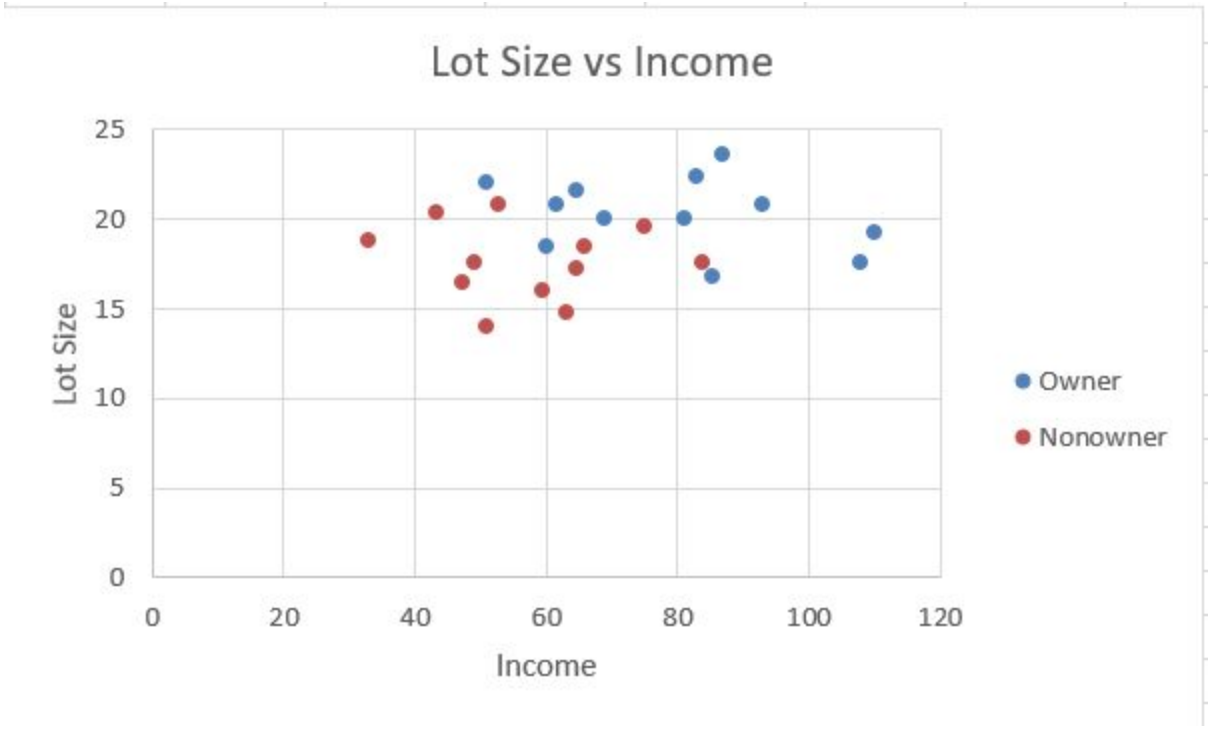


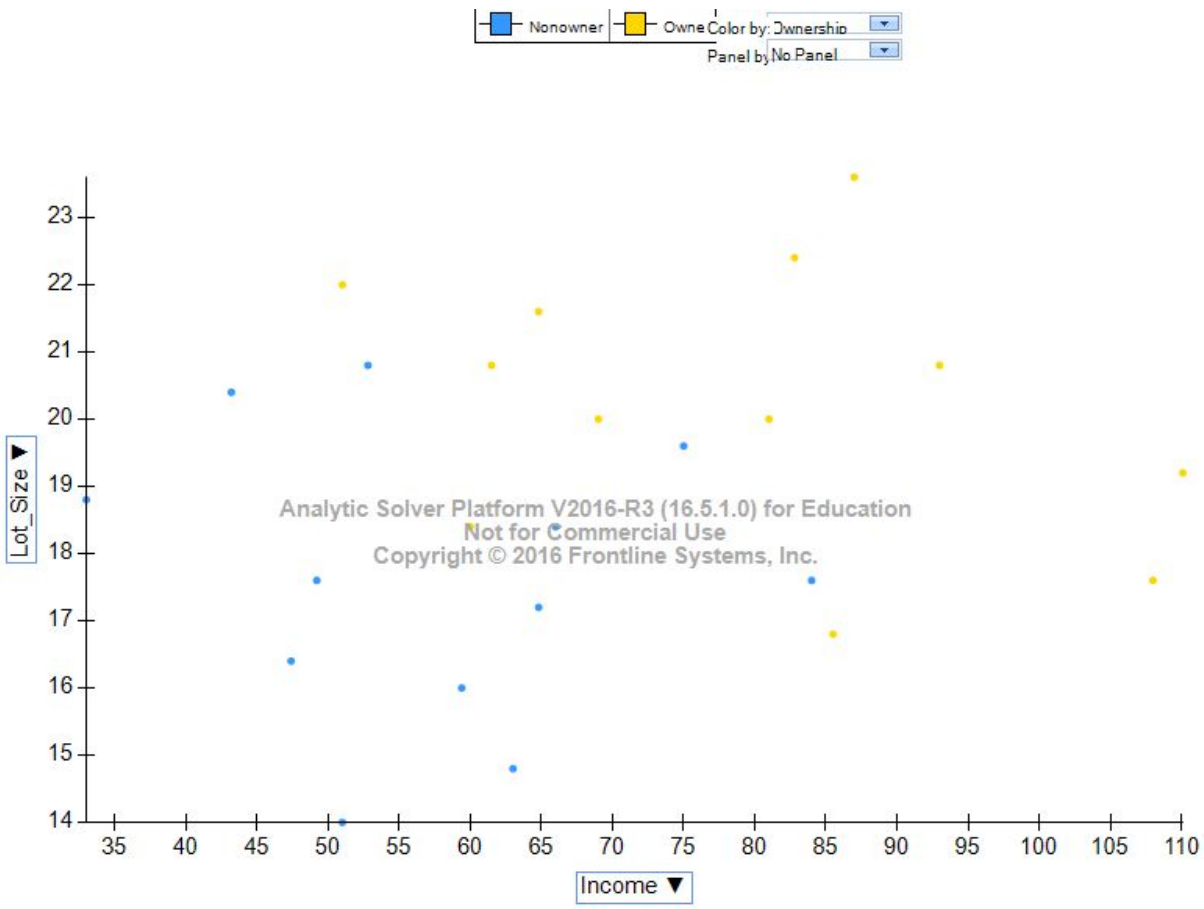
f. Microsoft Excel and XLMiner were quite similar in terms of the processes of generating the line graphs. After creating the final product, the quality of the graphs from both software were more or less the same. However, Microsoft Excel did prove to be easier to use when it came to creating the actual graphs. For example, when creating the multi-series time plot in Excel, it was really easy to just add one series, and then continue

adding series onto the graph. XLMiner also had some benefits as well. In XLMiner, if one accidentally plots the wrong variable on the wrong axis, it is really simple to change the variables on the axes using the drop-down menus.

Problem 3.2 15 points (see 3.2_RidingMowers.xlsx for data set)

a. Scatter plot of Lot Size vs. Income



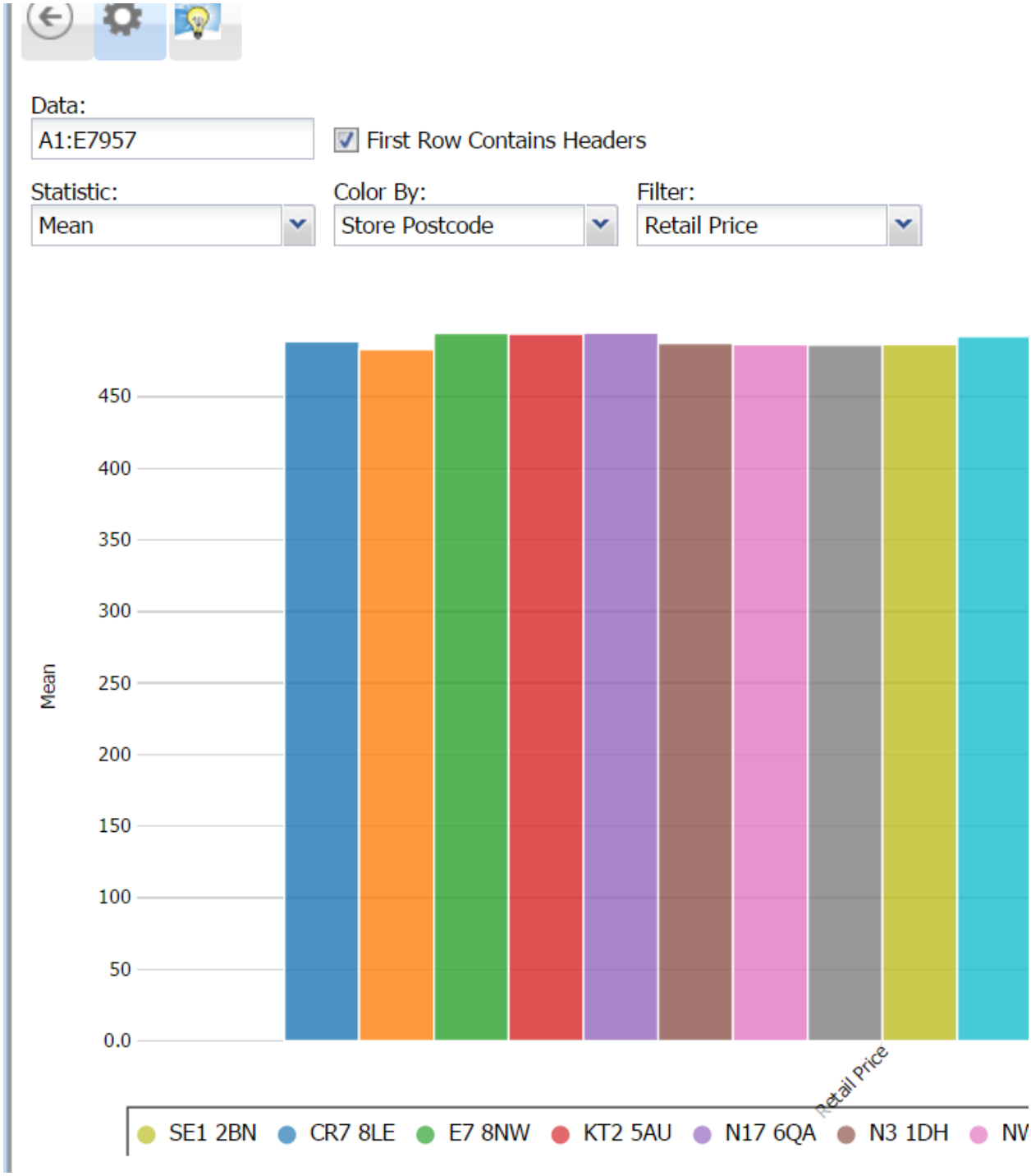


b.

c. Using Excel and XL Miner to create the identical plots was pretty straightforward in both cases. I will say that XL Miner was in fact slightly easier to create the plot, because in the plot creation, one can select a third variable to color by easily. In Excel, one has to add the data separately, once for each category that one would want to color by.

Problem 3.3 10 points (see 3.3_LaptopSalesJanuary2008.xlsx for data set)

a. Based on the visualization below, it appears that the store with the post code N17 6QA had the highest average, and E2 0RY had the lowest average. Please note that for some reason we could not zoom on these visualizations, so the actual highest or lowest might be different and we wrote down the wrong one.



b. Based on the side by side box plot, it appears that the two stores have quite similar price distributions. However, N17 does appear to be a little higher, and E2 appears to have a couple more outliers (N17 has a tighter distribution).

Problem 4.1 15 points (see 4.1_Cereals.xlsx for data set)

a. Quantitative/numerical: calories, protein, fat, sodium, fiber, carbo, sugars, potass, vitamins, shelf, weight, cups, rating

Ordinal: none

Nominal: name, mfr, type

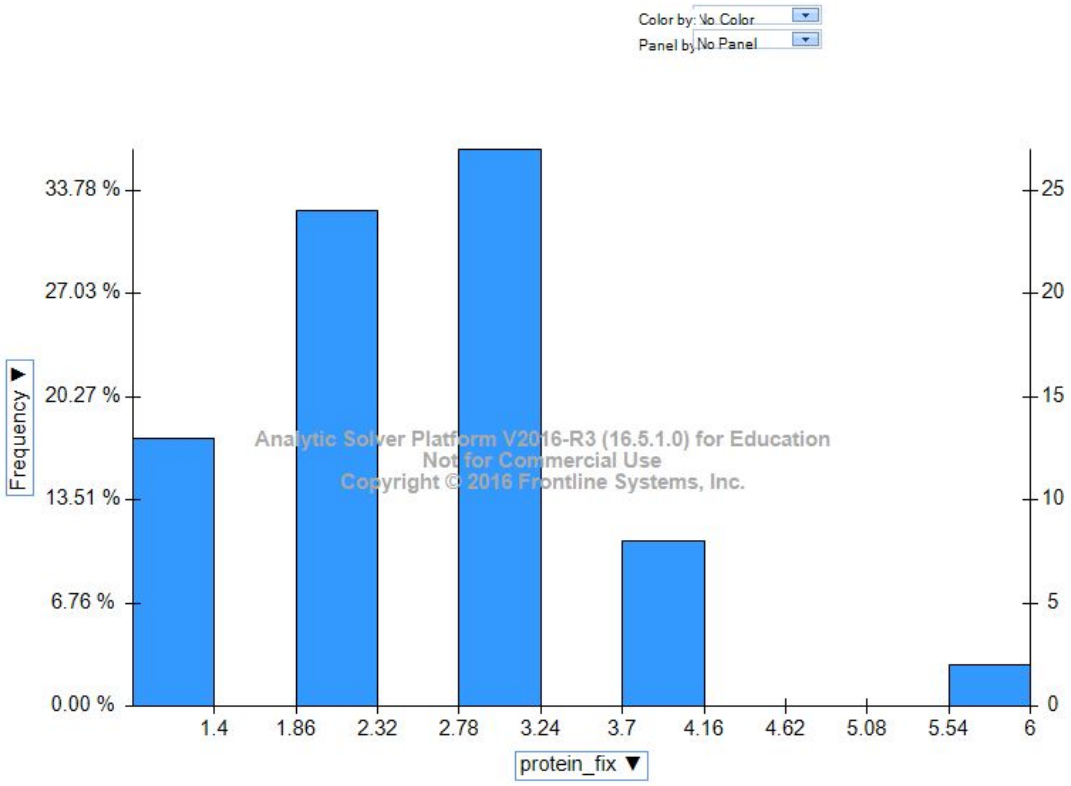
b. Descriptive Statistics for quantitative variables

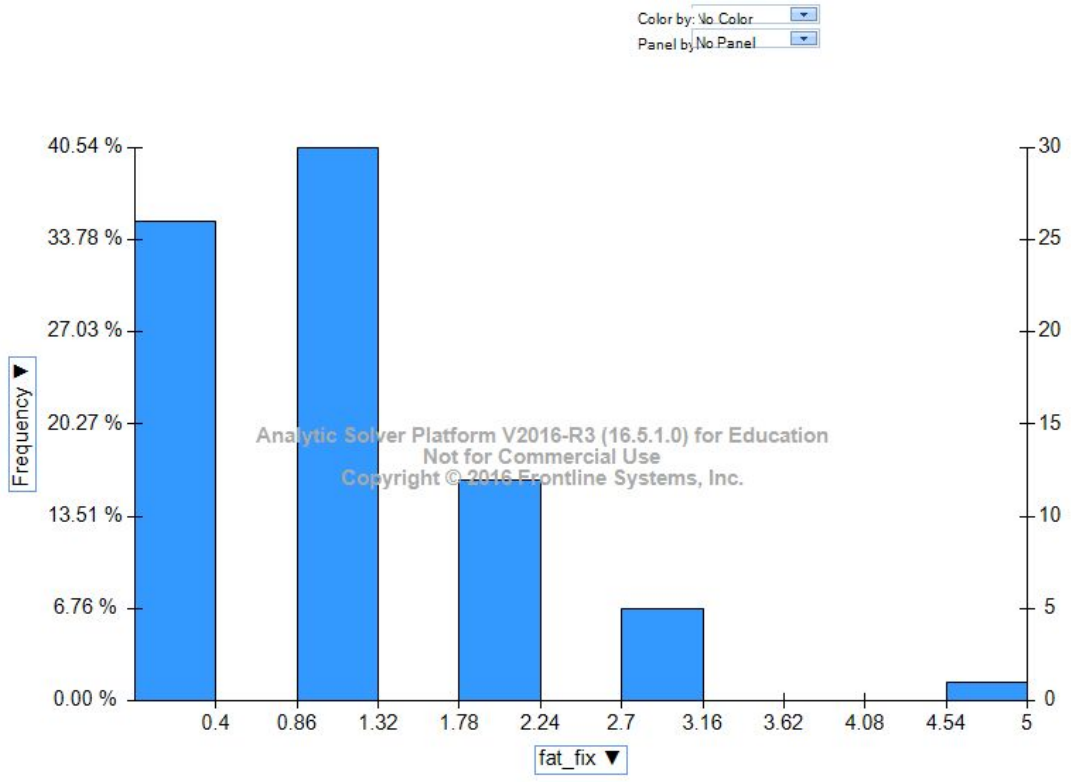
calories		*protein*		*fat*		*sodium*	
Mean	107.027027	Mean	2.51351351	Mean	1	Mean	162.364865
Standard Erro	2.30680568	Standard Erro	0.1250594	Standard Erro	0.11704115	Standard Erro	9.62179231
Median	110	Median	2.5	Median	1	Median	180
Mode	110	Mode	3	Mode	1	Mode	0
Standard Devi	19.8438928	Standard Devi	1.07580162	Standard Devi	1.00682602	Standard Devi	82.7697871
Sample Variar	393.780081	Sample Variar	1.15734913	Sample Variar	1.01369863	Sample Variar	6850.83765
Kurtosis	2.21775158	Kurtosis	1.38029107	Kurtosis	2.28559193	Kurtosis	-0.2175828
Skewness	-0.4606597	Skewness	0.74433527	Skewness	1.24152533	Skewness	-0.6116384
Range	110	Range	5	Range	5	Range	320
Minimum	50	Minimum	1	Minimum	0	Minimum	0
Maximum	160	Maximum	6	Maximum	5	Maximum	320
Sum	7920	Sum	186	Sum	74	Sum	12015
Count	74	Count	74	Count	74	Count	74

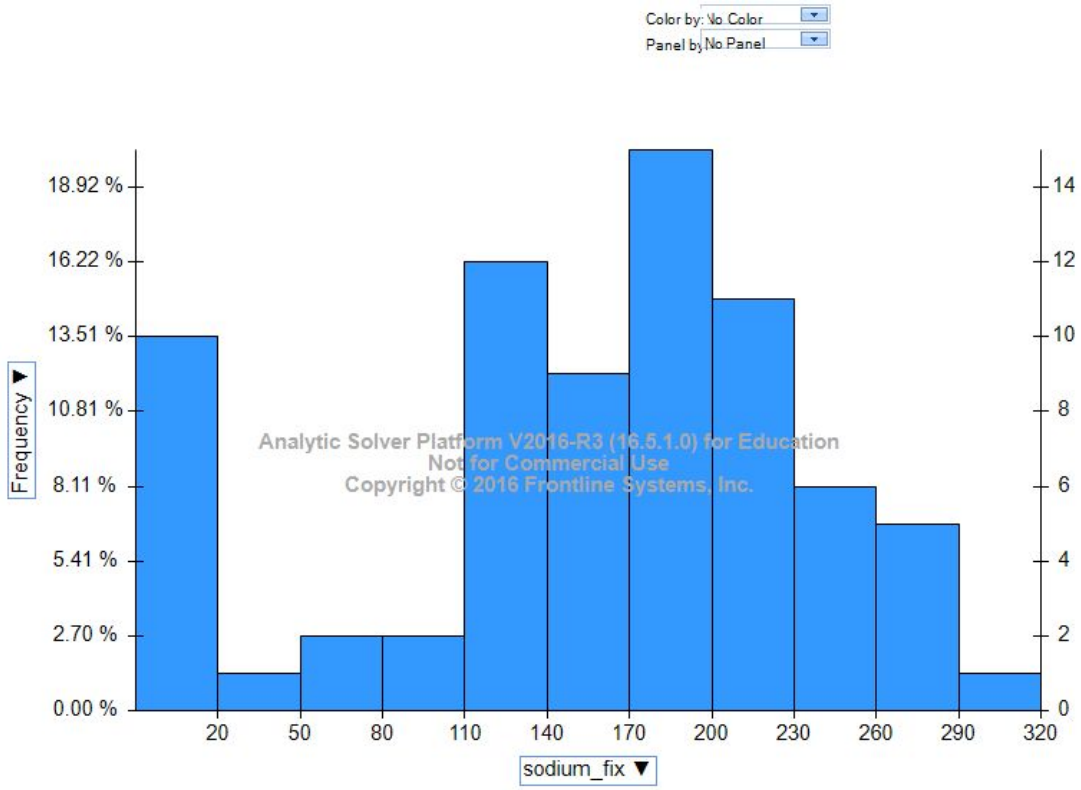
fiber		*carbo*		*sugars*		*potass*	
Mean	2.17567568	Mean	14.7297297	Mean	7.10810811	Mean	98.5135135
Standard Erro	0.2817135	Standard Erro	0.45239799	Standard Erro	0.50673639	Standard Erro	8.23947936
Median	2	Median	14.5	Median	7	Median	90
Mode	0	Mode	15	Mode	3	Mode	35
Standard Devi	2.42339119	Standard Devi	3.89167463	Standard Devi	4.35911128	Standard Devi	70.8786815
Sample Variar	5.87282488	Sample Variar	15.1451314	Sample Variar	19.0018512	Sample Variar	5023.78749
Kurtosis	8.27275697	Kurtosis	-0.3002671	Kurtosis	-1.168268	Kurtosis	1.91515575
Skewness	2.38350118	Skewness	0.11833498	Skewness	0.04725699	Skewness	1.39872172
Range	14	Range	18	Range	15	Range	315
Minimum	0	Minimum	5	Minimum	0	Minimum	15
Maximum	14	Maximum	23	Maximum	15	Maximum	330
Sum	161	Sum	1090	Sum	526	Sum	7290
Count	74	Count	74	Count	74	Count	74

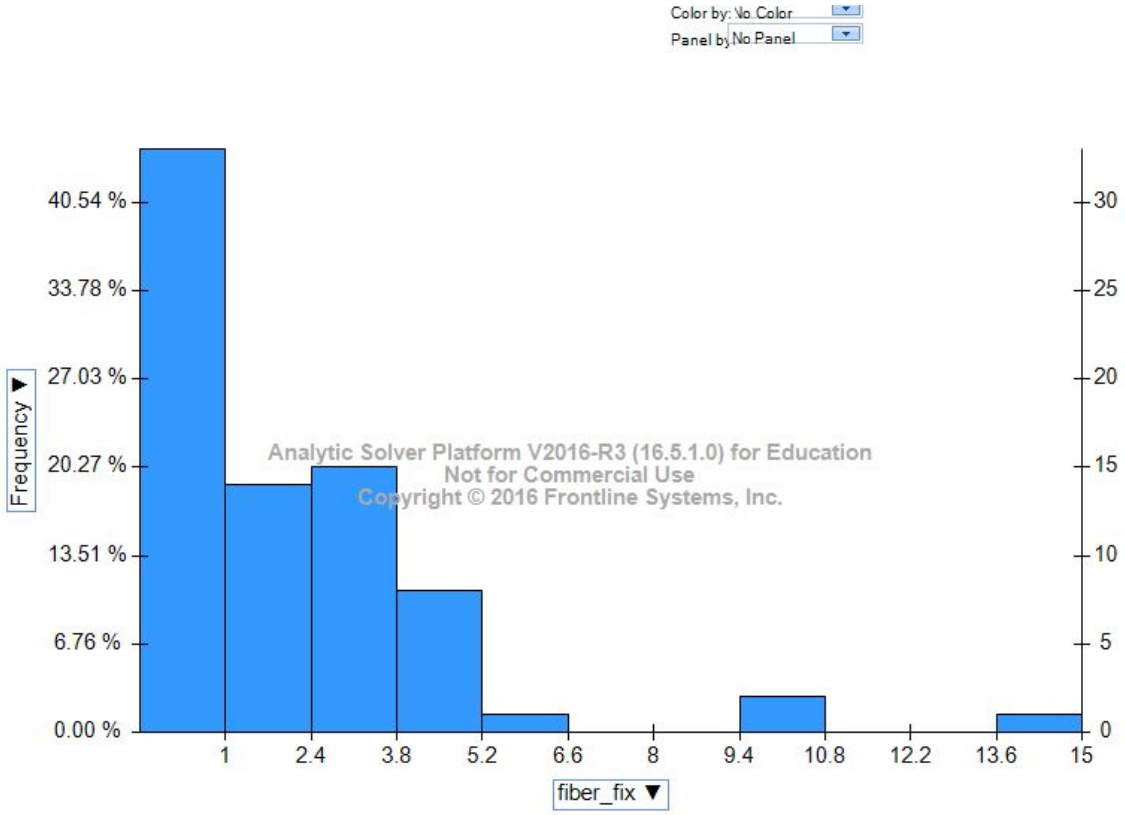
vitamins		*shelf*		*weight*		*cups*		*rating*	
Mean	29.0540541	Mean	2.21621622	Mean	1.03081081	Mean	0.82162162	Mean	42.3717869
Standard Erro	2.59166578	Standard Erro	0.09672587	Standard Erro	0.01783419	Standard Erro	0.02740135	Standard Erro	1.63138594
Median	25	Median	2	Median	1	Median	0.75	Median	40.2530865
Mode	25	Mode	3	Mode	1	Mode	1	Mode	#N/A
Standard Devi	22.2943521	Standard Devi	0.83206741	Standard Devi	0.15341551	Standard Devi	0.23571533	Standard Devi	14.0337125
Sample Variar	497.038134	Sample Variar	0.69233617	Sample Variar	0.02353632	Sample Variar	0.05556172	Sample Variar	196.945086
Kurtosis	6.28189286	Kurtosis	-1.426308	Kurtosis	5.12317255	Kurtosis	0.31397425	Kurtosis	1.52023145
Skewness	2.53002119	Skewness	-0.4284584	Skewness	0.28091207	Skewness	-0.1149834	Skewness	0.96286585
Range	100	Range	2	Range	1	Range	1.25	Range	75.662061
Minimum	0	Minimum	1	Minimum	0.5	Minimum	0.25	Minimum	18.042851
Maximum	100	Maximum	3	Maximum	1.5	Maximum	1.5	Maximum	93.704912
Sum	2150	Sum	164	Sum	76.28	Sum	60.8	Sum	3135.51223
Count	74	Count	74	Count	74	Count	74	Count	74

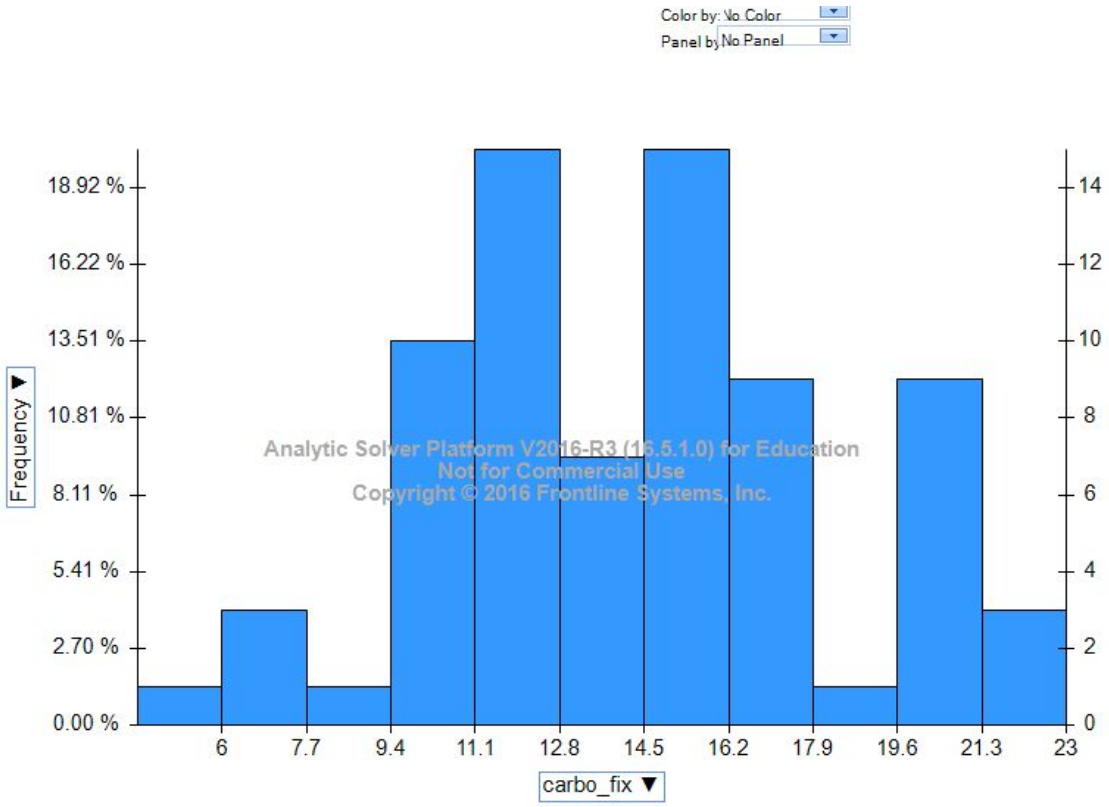
c. Below are the histogram screen shots

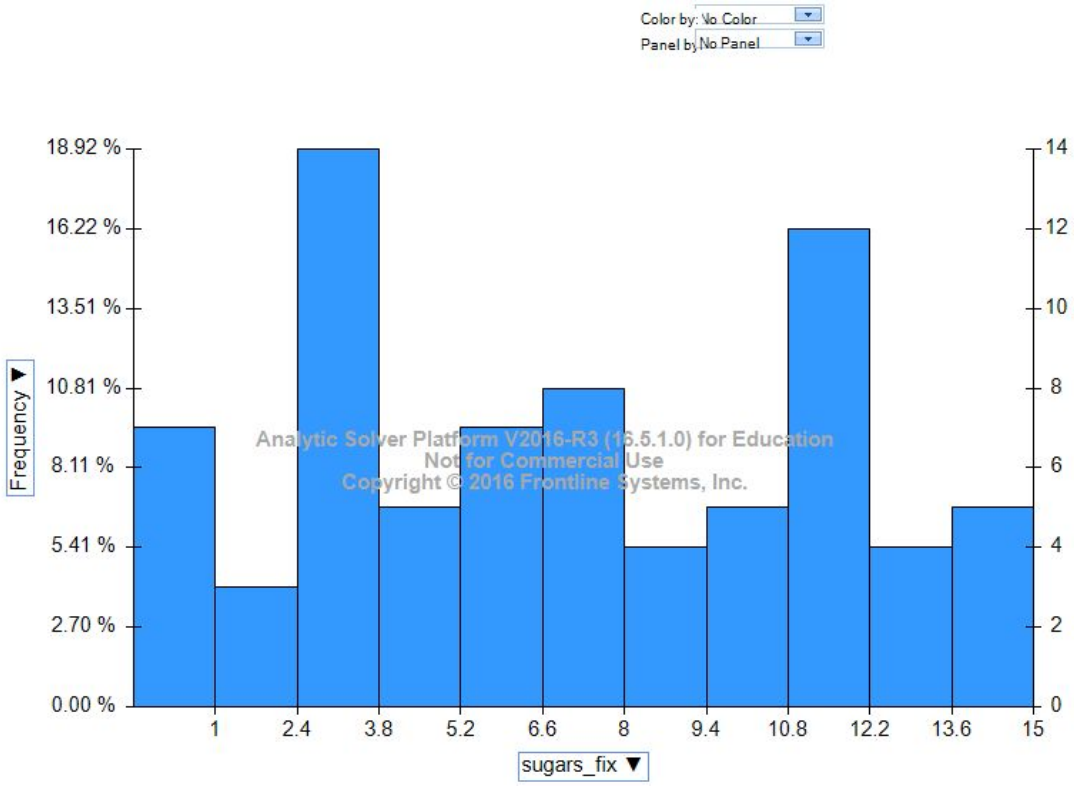

Color by: No Color
Panel by: No Panel
33.78 %
27.03 %
20.27 %
13.51 %
6.76 %
0.00 %
Frequency
25
20
15
10
5
0
1.4
1.86
2.32
2.78
3.24
3.7
4.16
4.62
5.08
5.54
6
protein_fix
Analytic Solver Platform V2016-R3 (16.5.1.0) for Education
Not for Commercial Use
Copyright © 2016 Frontline Systems, Inc.

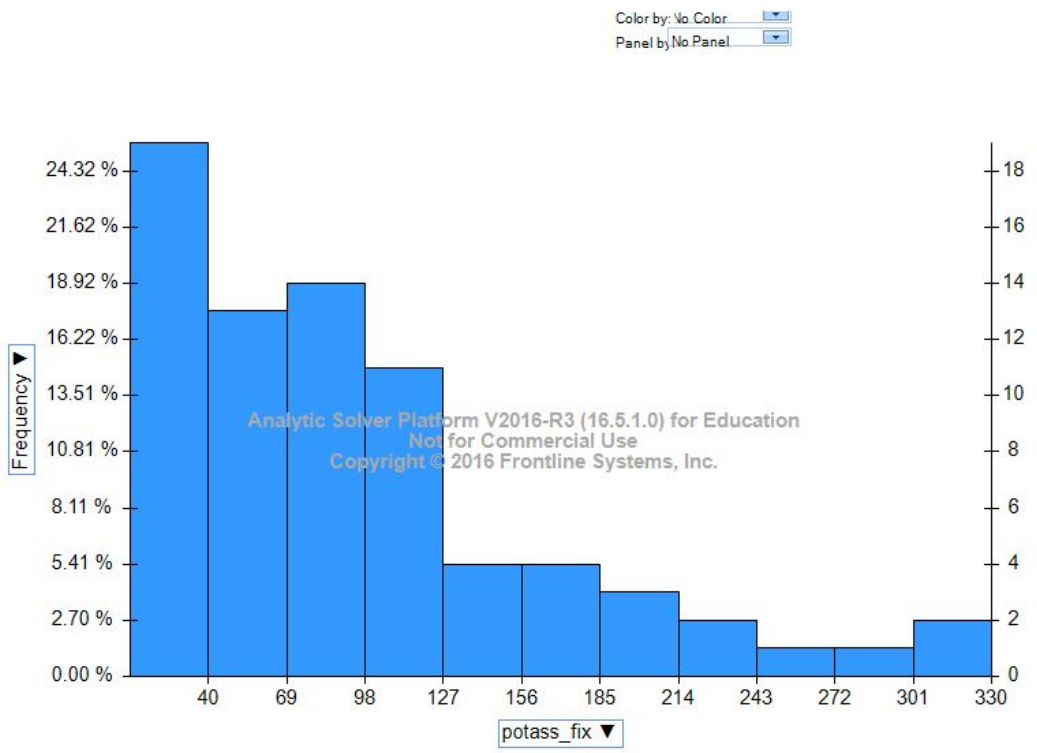

Color by: No Color
Panel by: No Panel
40.54 %
33.78 %
27.03 %
20.27 %
13.51 %
6.76 %
0.00 %
Frequency
30
25
20
15
10
5
0
0.4
0.86
1.32
1.78
2.24
2.7
3.16
3.62
4.08
4.54
5
fat_fix
Analytic Solver Platform V2016-R3 (16.5.1.0) for Education
Not for Commercial Use
Copyright © 2016 Frontline Systems, Inc.

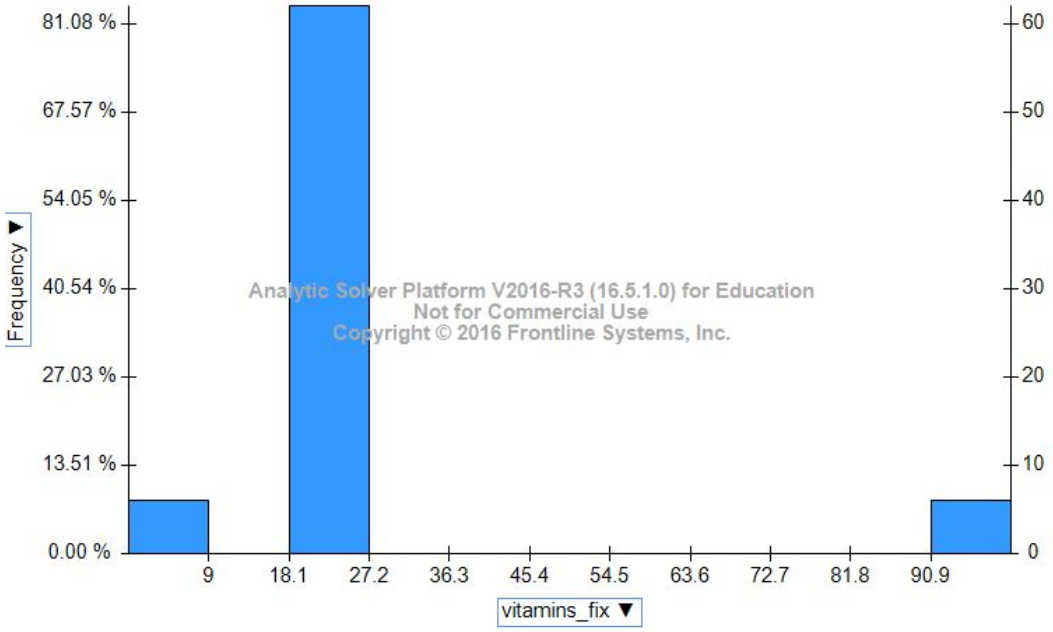

Color by: No Color
Panel by: No Panel
18.92 %
16.22 %
13.51 %
10.81 %
8.11 %
5.41 %
2.70 %
0.00 %
Frequency
14
12
10
8
6
4
2
0
20
50
80
110
140
170
200
230
260
290
320
sodium_fix
Analytic Solver Platform V2016-R3 (16.5.1.0) for Education
Not for Commercial Use
Copyright © 2016 Frontline Systems, Inc.

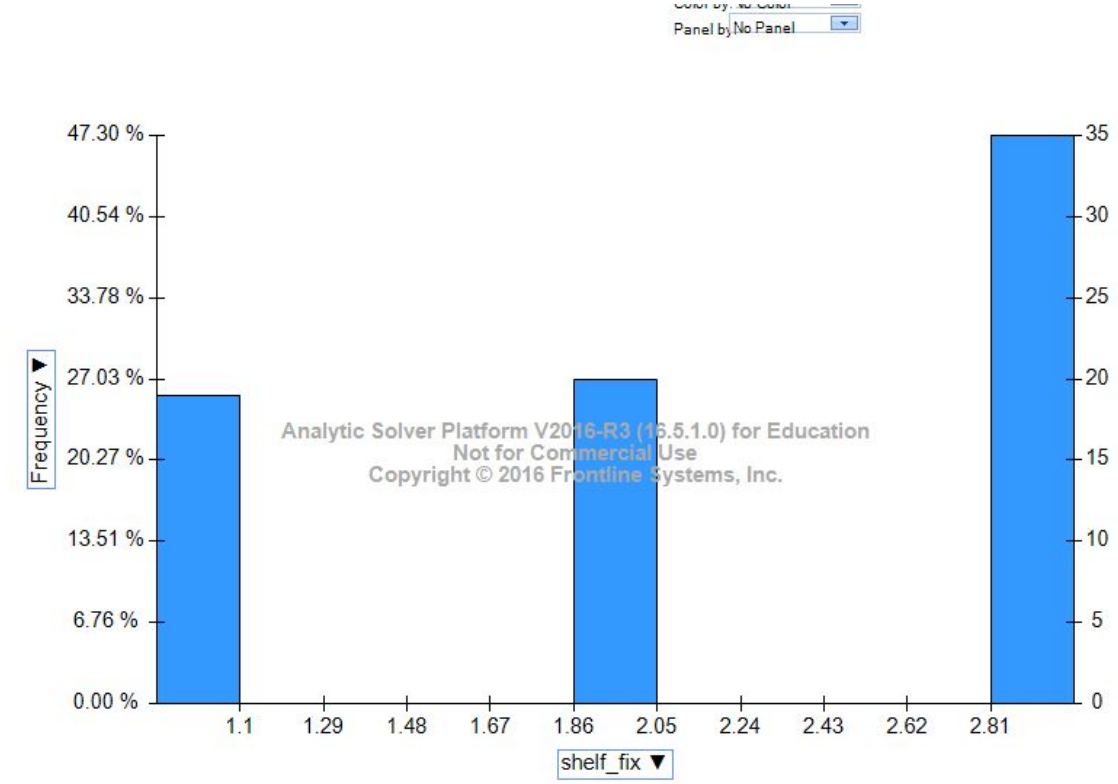

Color by: No Color
Panel by: No Panel
40.54 %
33.78 %
27.03 %
20.27 %
13.51 %
6.76 %
0.00 %
Frequency
30
25
20
15
10
5
0
1
2.4
3.8
5.2
6.6
8
9.4
10.8
12.2
13.6
15
fiber_fix
Analytic Solver Platform V2016-R3 (16.5.1.0) for Education
Not for Commercial Use
Copyright © 2016 Frontline Systems, Inc.

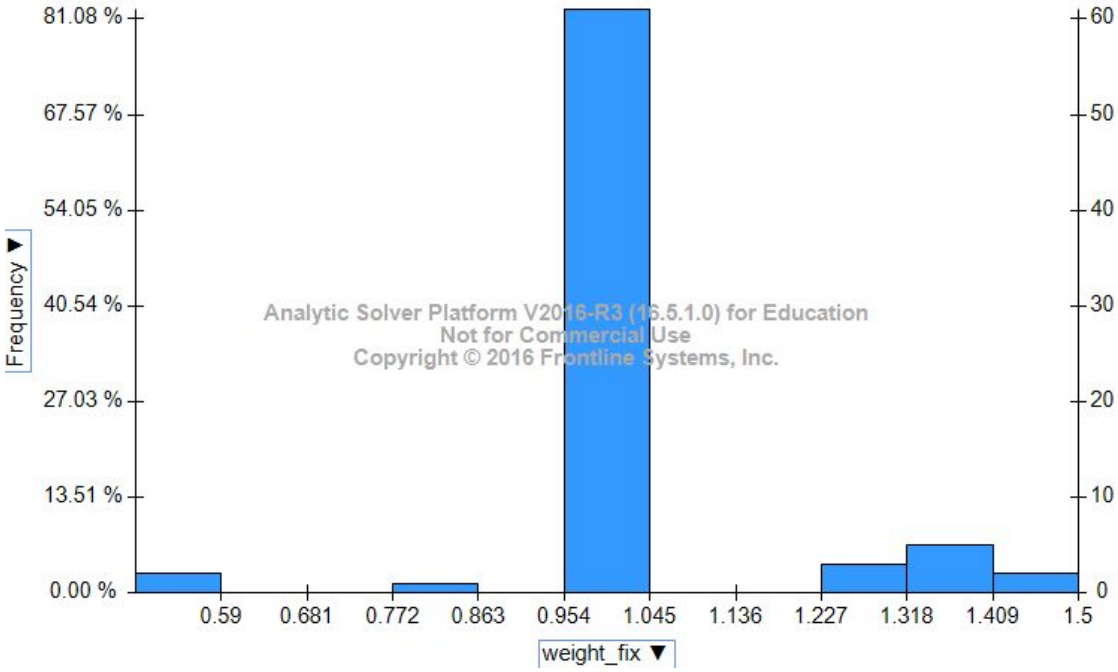

Color by: No Color
Panel by: No Panel
18.92 %
16.22 %
13.51 %
10.81 %
8.11 %
5.41 %
2.70 %
0.00 %
Frequency
14
12
10
8
6
4
2
0
6
7.7
9.4
11.1
12.8
14.5
16.2
17.9
19.6
21.3
23
carbo_fix
Analytic Solver Platform V2016-R3 (16.5.1.0) for Education
Not for Commercial Use
Copyright © 2016 Frontline Systems, Inc.

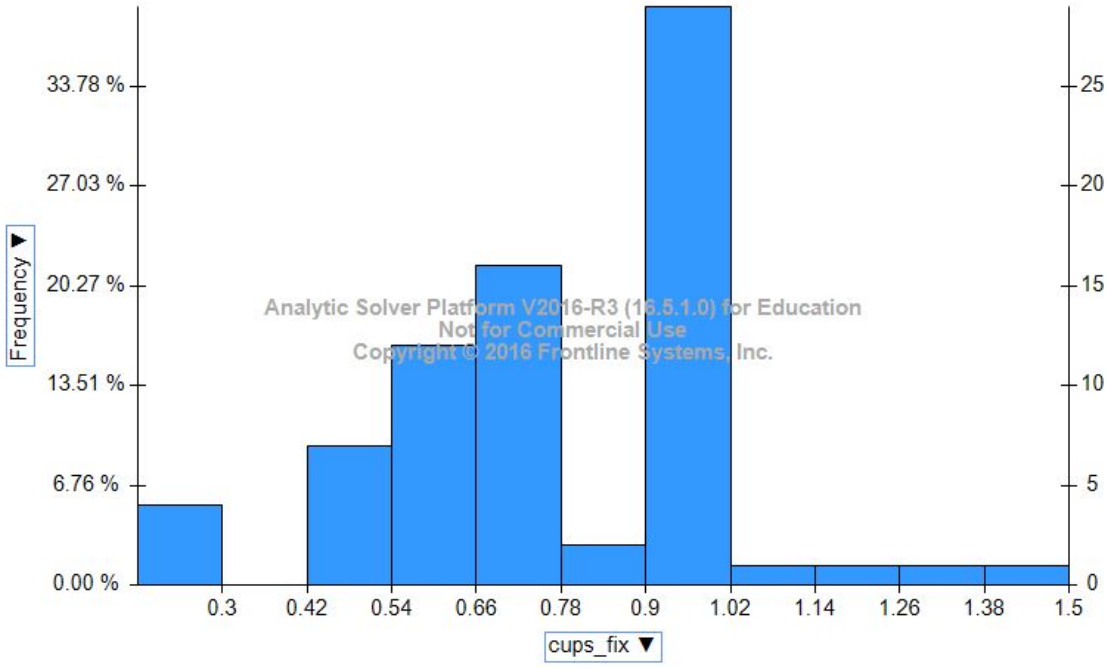

Color by: No Color
Panel by: No Panel
18.92 %
16.22 %
13.51 %
10.81 %
8.11 %
5.41 %
2.70 %
0.00 %
Frequency
14
12
10
8
6
4
2
0
1
2.4
3.8
5.2
6.6
8
9.4
10.8
12.2
13.6
15
sugars_fix
Analytic Solver Platform V2016-R3 (16.5.1.0) for Education
Not for Commercial Use
Copyright © 2016 Frontline Systems, Inc.

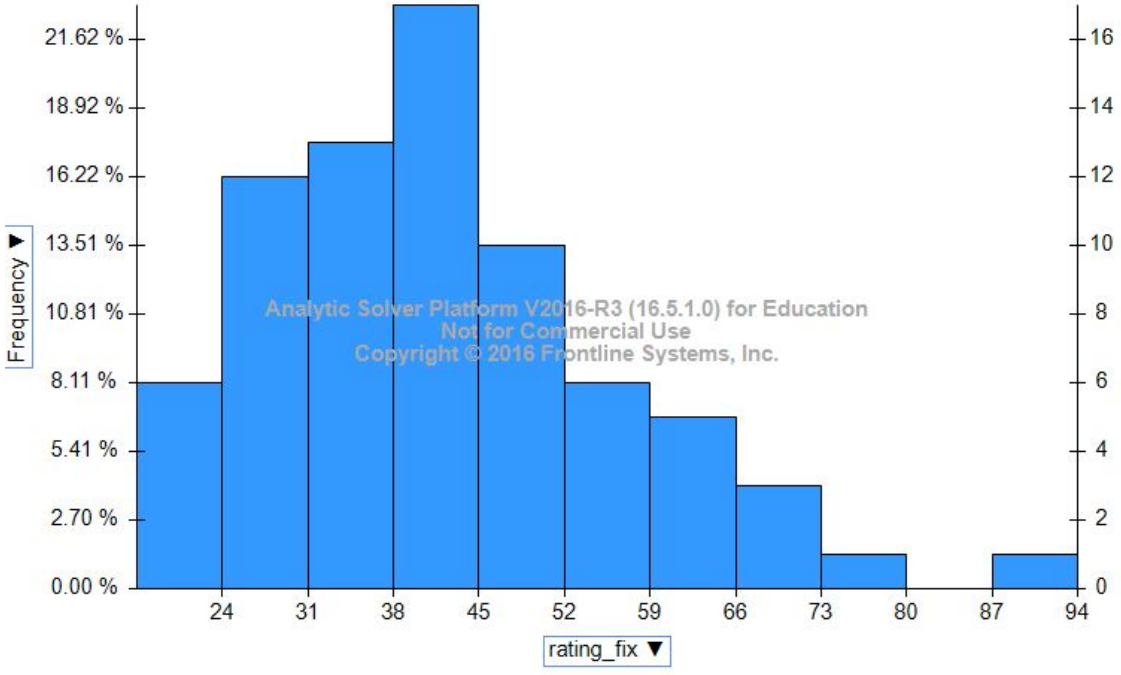

Color by: No Color
Panel by: No Panel
24.32 %
21.62 %
18.92 %
16.22 %
13.51 %
10.81 %
8.11 %
5.41 %
2.70 %
0.00 %
Frequency
18
16
14
12
10
8
6
4
2
0
40
69
98
127
156
185
214
243
272
301
330
potass_fix
Analytic Solver Platform V2016-R3 (16.5.1.0) for Education
Not for Commercial Use
Copyright © 2016 Frontline Systems, Inc.


81.08 %
67.57 %
54.05 %
40.54 %
27.03 %
13.51 %
0.00 %
Frequency
60
50
40
30
20
10
0
9
18.1
27.2
36.3
45.4
54.5
63.6
72.7
81.8
90.9
vitamins_fix
Analytic Solver Platform V2016-R3 (16.5.1.0) for Education
Not for Commercial Use
Copyright © 2016 Frontline Systems, Inc.

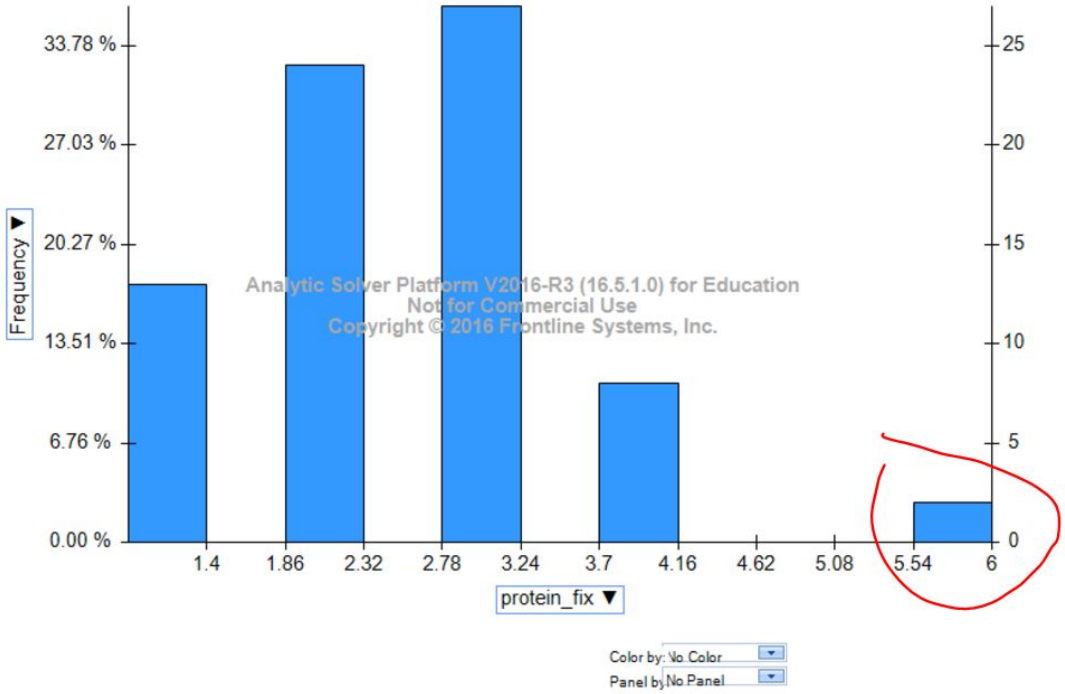

Panel by No Panel
47.30 %
40.54 %
33.78 %
27.03 %
20.27 %
13.51 %
6.76 %
0.00 %
Frequency
35
30
25
20
15
10
5
0
1.1
1.29
1.48
1.67
1.86
2.05
2.24
2.43
2.62
2.81
shelf_fix
Analytic Solver Platform V2016-R3 (16.5.1.0) for Education
Not for Commercial Use
Copyright © 2016 Frontline Systems, Inc.

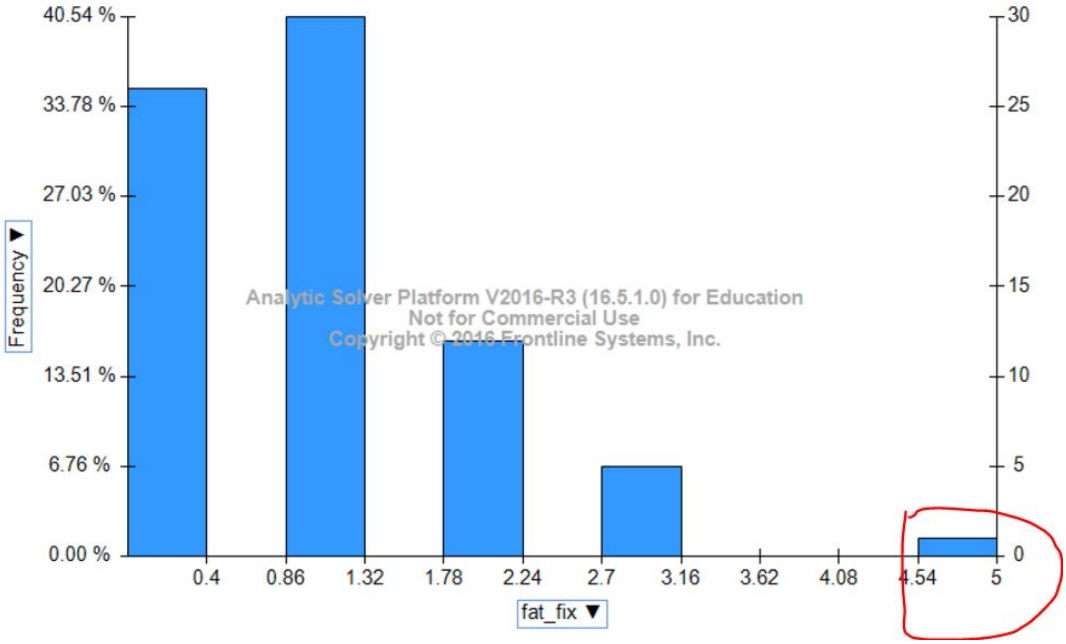

81.08 %
67.57 %
54.05 %
40.54 %
27.03 %
13.51 %
0.00 %
Frequency
60
50
40
30
20
10
0
0.59
0.681
0.772
0.863
0.954
1.045
1.136
1.227
1.318
1.409
1.5
weight_fix
Analytic Solver Platform V2016-R3 (16.5.1.0) for Education
Not for Commercial Use
Copyright © 2016 Frontline Systems, Inc.

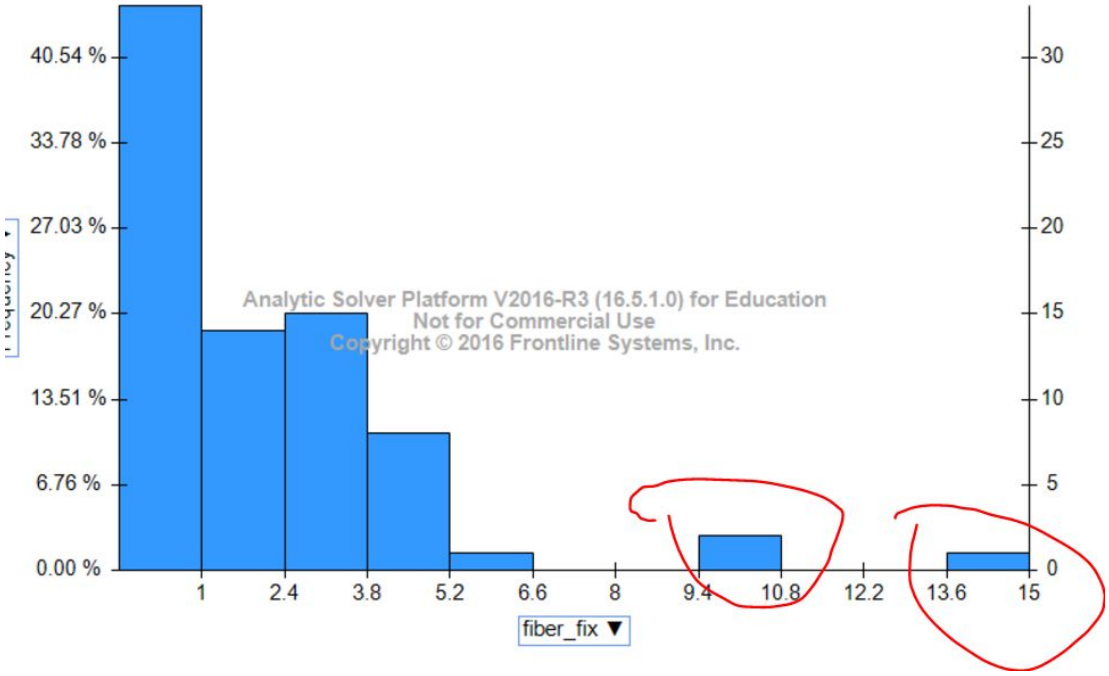

33.78 %
27.03 %
20.27 %
13.51 %
6.76 %
0.00 %
Frequency
25
20
15
10
5
0
0.3
0.42
0.54
0.66
0.78
0.9
1.02
1.14
1.26
1.38
1.5
cups_fix
Analytic Solver Platform V2016-R3 (16.5.1.0) for Education
Not for Commercial Use
Copyright © 2016 Frontline Systems, Inc.

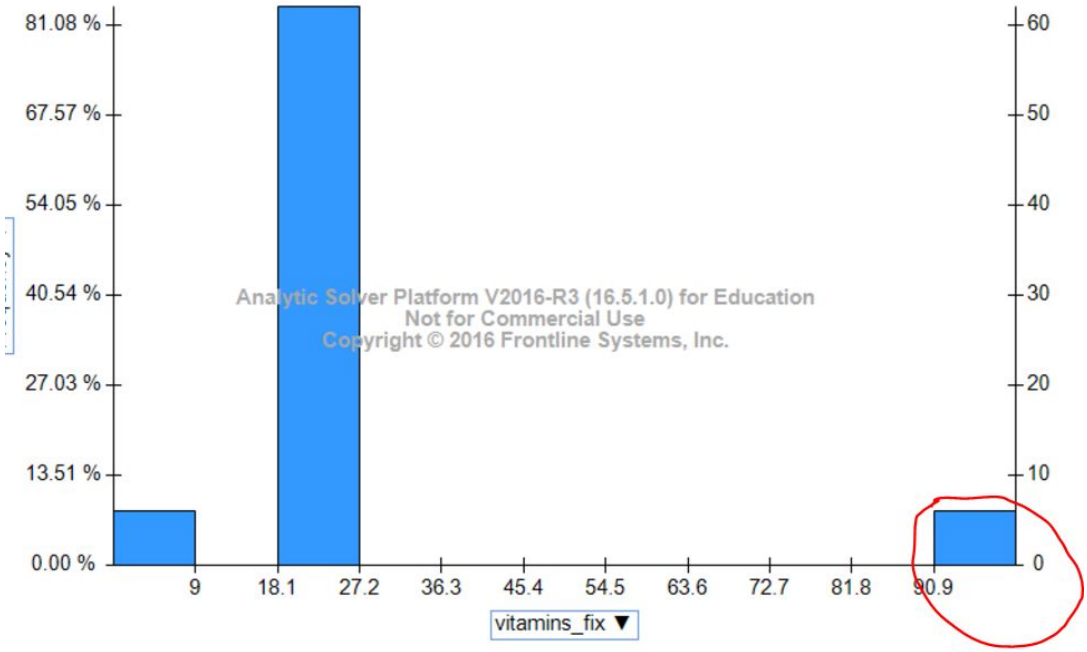

21.62 %
18.92 %
16.22 %
13.51 %
10.81 %
8.11 %
5.41 %
2.70 %
0.00 %
Frequency
16
14
12
10
8
6
4
2
0
24
31
38
45
52
59
66
73
80
87
94
rating_fix
Analytic Solver Platform V2016-R3 (16.5.1.0) for Education
Not for Commercial Use
Copyright © 2016 Frontline Systems, Inc.

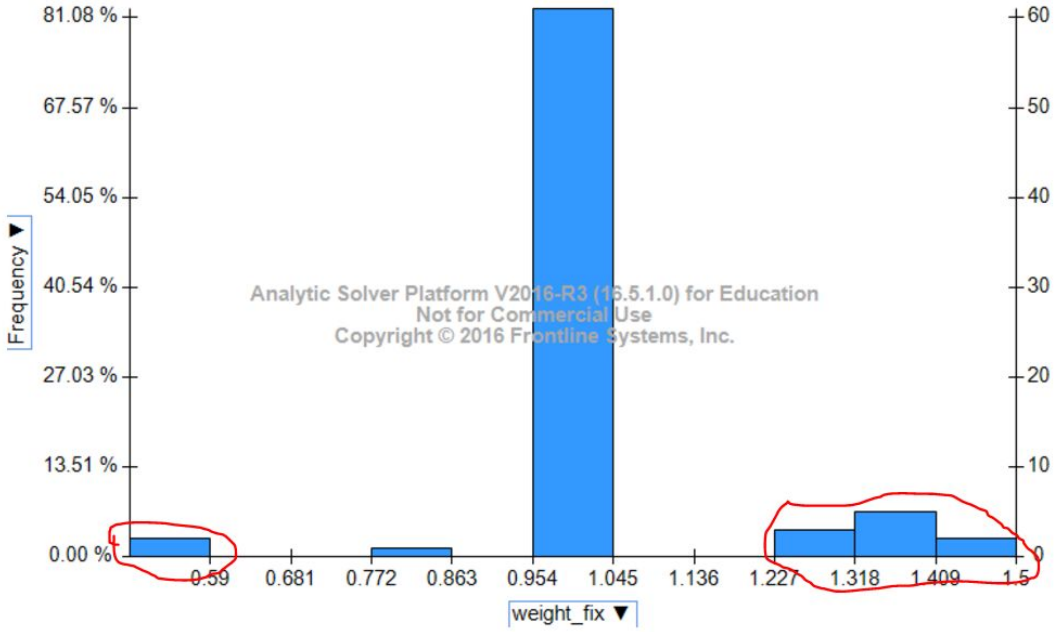
i. Using standard deviation to measure variability, the variables calories, sugar, sodium, and potassium were the variables with the largest variability.
ii. Rating, vitamins, potassium, fat and fiber all appear to to be quite skewed (either left or right).
iii. In several of the histograms, there appear to be some extreme values. Below are the histograms from above, except with potentially extreme values circled in red.

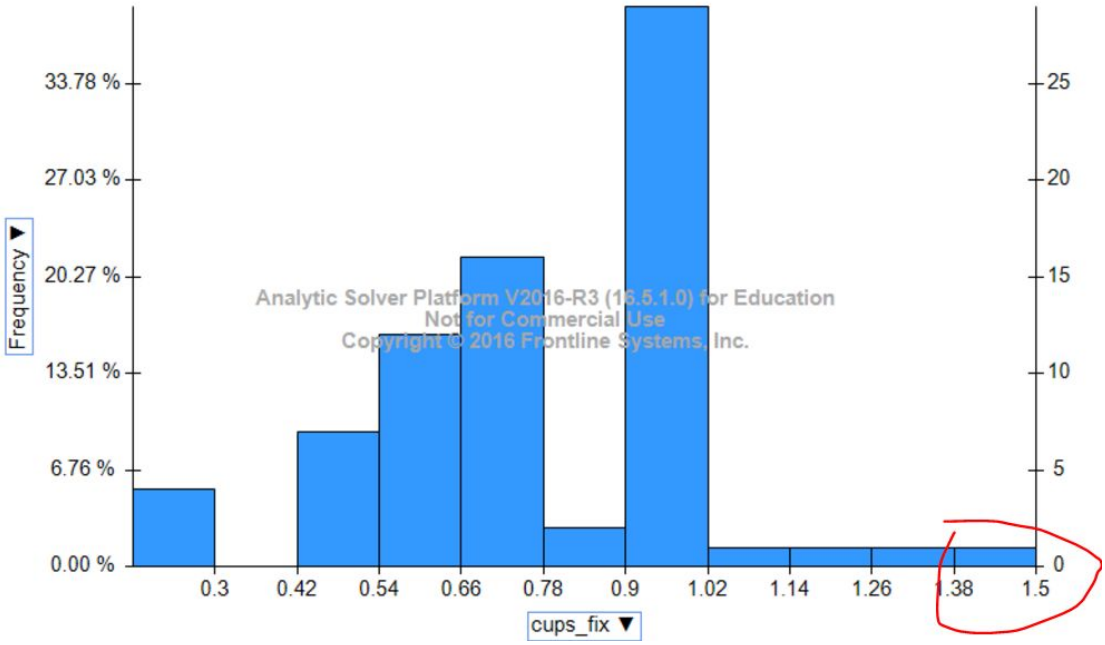


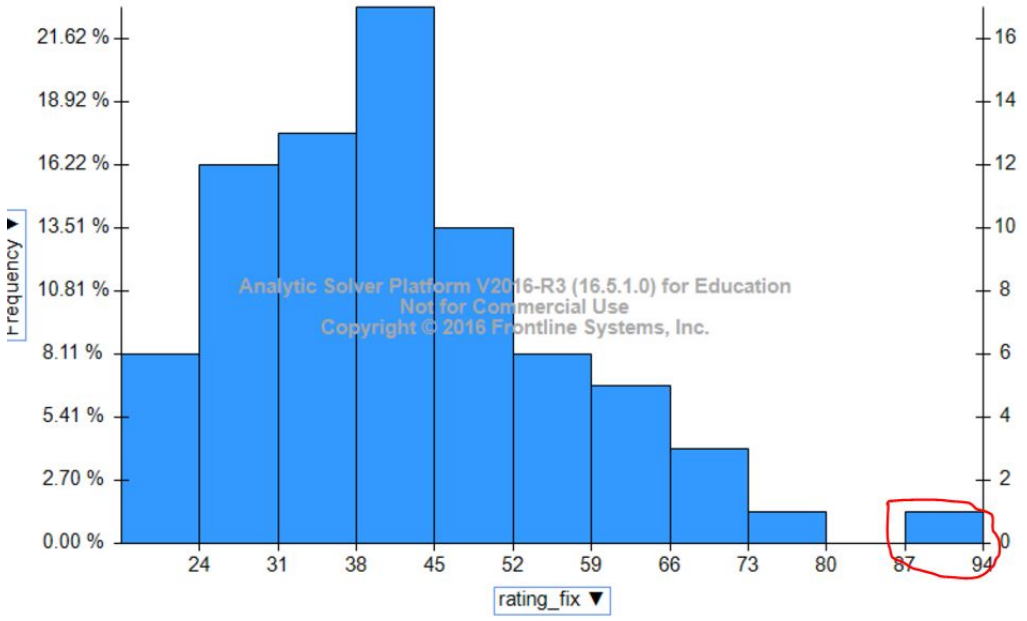


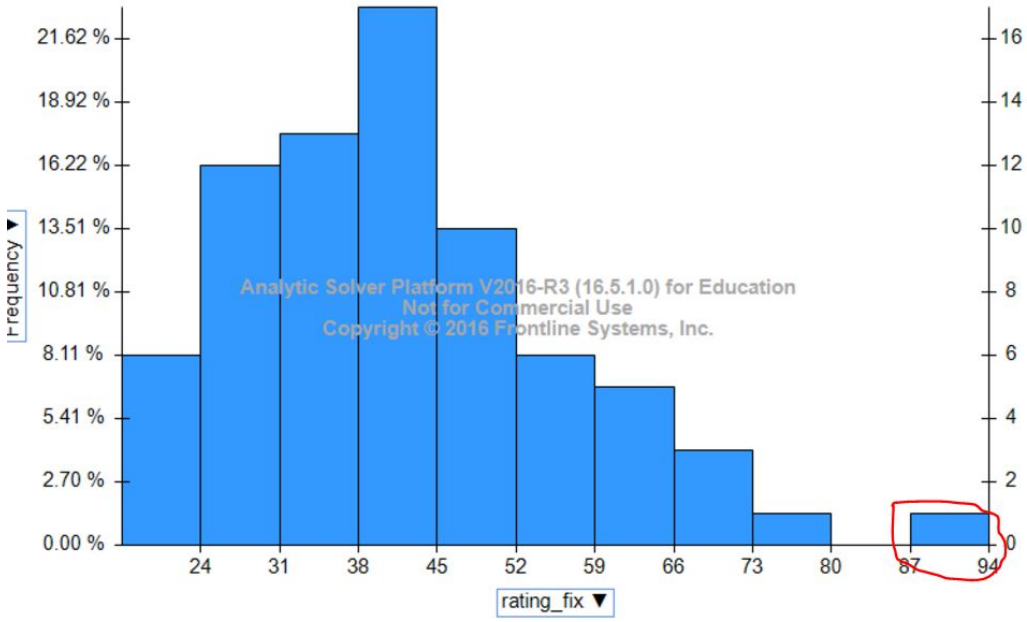

40.54 %
33.78 %
27.03 %
20.27 %
13.51 %
6.76 %
0.00 %
30
25
20
15
10
5
0
1
2.4
3.8
5.2
6.6
8
9.4
10.8
12.2
13.6
15
fiber_fix ▼
Analytic Solver Platform V2016-R3 (16.5.1.0) for Education
Not for Commercial Use
Copyright © 2016 Frontline Systems, Inc.


81.08 %
67.57 %
54.05 %
40.54 %
27.03 %
13.51 %
0.00 %
60
50
40
30
20
10
0
9
18.1
27.2
36.3
45.4
54.5
63.6
72.7
81.8
90.9
vitamins_fix ▼
Analytic Solver Platform V2016-R3 (16.5.1.0) for Education
Not for Commercial Use
Copyright © 2016 Frontline Systems, Inc.

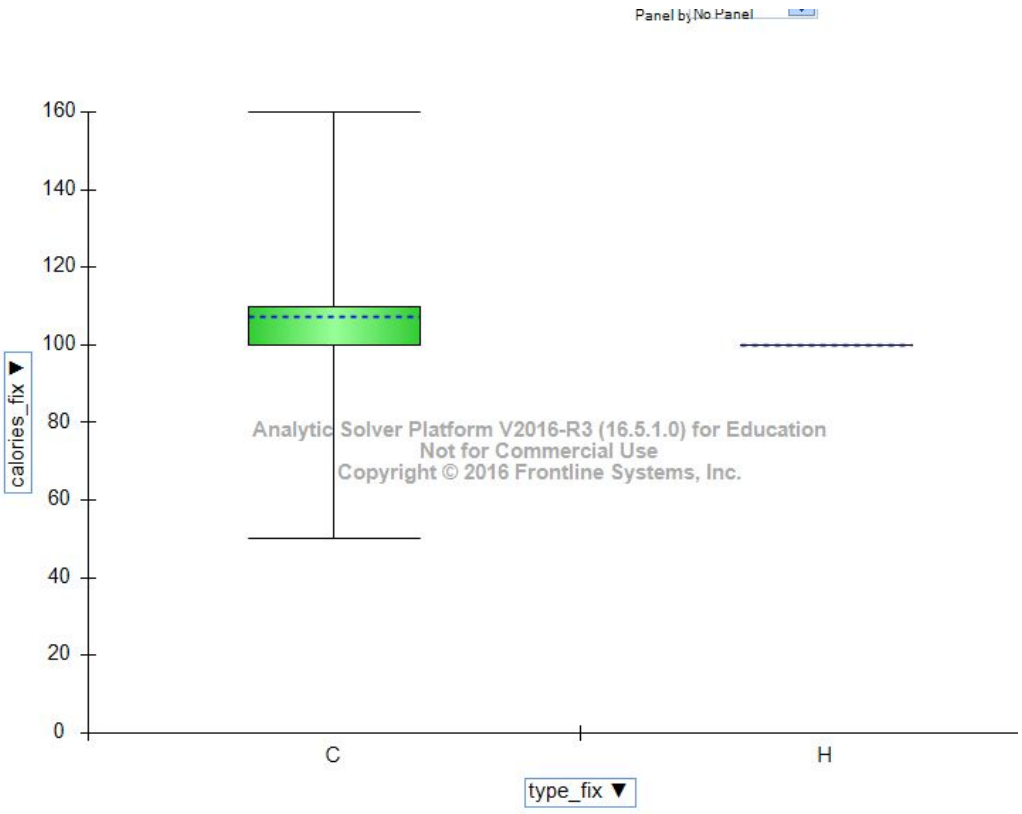

81.08 %
67.57 %
54.05 %
40.54 %
27.03 %
13.51 %
0.00 %
Frequency
60
50
40
30
20
10
0
0.59
0.681
0.772
0.863
0.954
1.045
1.136
1.227
1.318
1.409
1.5
weight_fix
Analytic Solver Platform V2016-R3 (16.5.1.0) for Education
Not for Commercial Use
Copyright © 2016 Frontline Systems, Inc.


33.78 %
27.03 %
20.27 %
13.51 %
6.76 %
0.00 %
Frequency
25
20
15
10
5
0
0.3
0.42
0.54
0.66
0.78
0.9
1.02
1.14
1.26
1.38
1.5
cups_fix
Analytic Solver Platform V2016-R3 (16.5.1.0) for Education
Not for Commercial Use
Copyright © 2016 Frontline Systems, Inc.

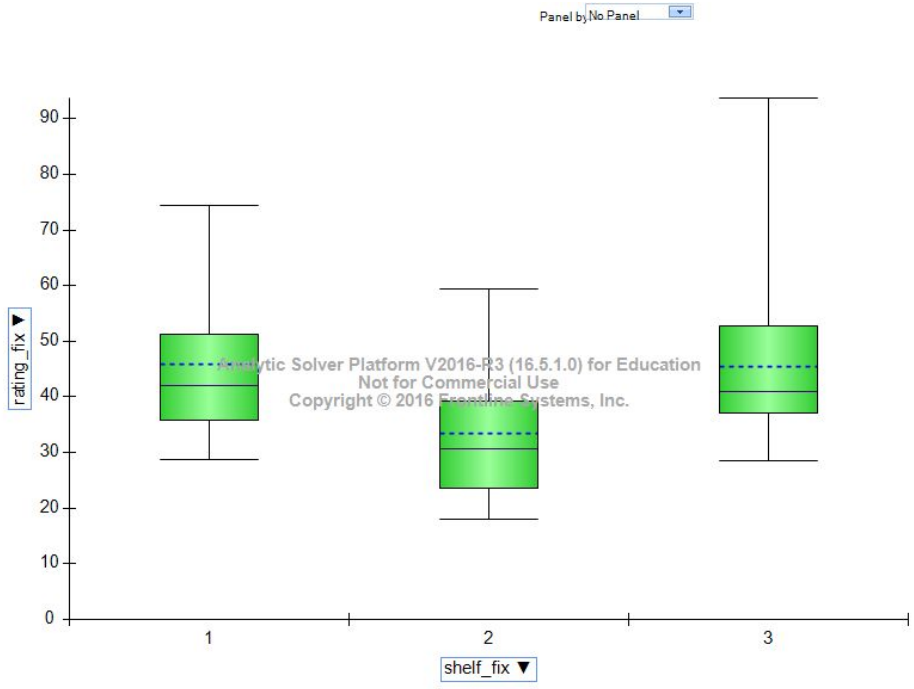




d. Based on the side by side box plot below, it appears that both hot and cold cereals tend to have the same amount of calories, which is around 100. However, for a more accurate analysis, more hot cereal data would probably be needed, as this sample only had one hot cereal in it.



e. As you can see in the side-by-side box plots below, shelf 1 and shelf 3 have significant overlap, and are therefore not statistically significant (they are telling us the same thing). We could remove one of the two categories of shelf height after consulting with the domain expert.



f. Correlation table

	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
calories	1												
protein	0.03399166	1											
fat	0.50737324	0.2023534	1										
sodium	0.2962475	0.01155889	0.0008219	1									
fiber	-0.2952118	0.5140061	0.01403587	-0.0707349	1								
carbo	0.27060605	-0.0367433	-0.2849337	0.32840919	-0.3790837	1							
sugars	0.56912054	-0.286584	0.28715249	0.03705896	-0.1509485	-0.4520692	1						
potass	-0.0713612	0.57874284	0.19963672	-0.0394381	0.91150392	-0.3650029	0.00141398	1					
vitamins	0.25984556	0.05479952	-0.0305139	0.33157596	-0.0387173	0.25357897	0.07295438	-0.0026358	1				
shelf	0.08924278	0.19563468	0.27797972	-0.1218968	0.31378736	-0.1889963	0.06144909	0.39458548	0.28440479	1			
weight	0.69645215	0.23067141	0.22171416	0.31253357	0.24629218	0.14480528	0.46054713	0.42056153	0.3204348	0.19284304	1		
cups	0.08919615	-0.2420986	-0.1575787	0.11958411	-0.5136972	0.35828371	-0.0324361	-0.5016883	0.13362965	-0.3510335	-0.2017146	1	
rating	-0.6937847	0.46716218	-0.4050502	-0.3830124	0.6034109	0.05594129	-0.7559551	0.41578244	-0.2144809	0.05103975	-0.300461	-0.2225044	1

i. Variables that are strongly correlated (assuming strong correlations are > 0.3 or < -0.3): fat and calories, sugars and calories, weight and calories, rating and calories, protein and fiber, protein and potassium, protein and rating, fat and rating, sodium and carbo, sodium and vitamins, sodium and weight, sodium and rating, fiber and carbo, fiber and potassium, fiber and shelf, fiber and cups, fiber and rating, carbo and sugars, carbo and potassium, carbo and cups, sugars and weight, sugars and rating, potassium and shelf, potassium and weight, potassium and cups, potassium and rating, vitamins and weight, shelf and cups, weight and rating

ii. We can reduce the variables by eliminating variables that have high correlations because one of the two highly correlated variables is giving us all the information necessary, and the second variable is technically redundant.

iii. Since all of the values in the dataset would be within the same range, it is possible that the correlations would be more accurate.

g. The first column on the left indicates that 86.319% of the variance in the original data is captured in the first principal component.

Principal Components		
Feature\Component	1	2
calories	–0.847053	–0.531508
rating	0.5315077	–0.847053

Variances		
	1	2
Variance	498.02448	78.932739
Variance %	86.319135	13.680865
Cumulative Variance %	86.319135	100

FIGURE 4.10 PCA OUTPUT USING ALL 13 NUMERICAL VARIABLES IN THE BREAKFAST CEREALS DATASET.

Problem 4.2 15 points (see 4.2_Wine.xlsx for data set)

a. In PCA, the data miner must order the principal components by variance. Therefore, the first principal component has the highest variance.

b. Normalizing the data prior to performing PCA would scale the values to give them all equal importance in terms of their variability. This would give the data miner more accurate variances for each principal component.

Problem 4.3 15 points (see 4.3_University.xlsx for data set)

a. We removed categorical data by deleting the columns with categorical variables, which were College Name and State. Then, we removed all missing data using the "Missing Data Handling" tool in XL Miner.

b. After performing the PCA, we get the results shown below. We can see that the first two principal components account for over 92% of the variance. The data should definitely be normalized before performing PCA to bring the values into the same numerical range and give their variabilities equal weight. For example, the variable graduation rate has values between 0 and 100, whereas tuition ranges from 1,000 to around 20,000 (therefore, tuition would dominate graduation rate). After standardizing the data, we found that the first 10 components now account for the same 92% of variance. The key principal components are those that account for the majority of the variance as they are best predictors.

Non-standardized:

Variances																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Variance	55218483	35856012	3439694	1422126.1	935916.79	461927.78	356380.48	337131.08	174401.15	101204.9064	35670.974	24213.281	362.86983	156.96945	121.40747	28.409239	8.445782	0.0286374
Variance I	56.136974	36.452432	3.4969091	1.4457814	0.9514846	0.4696114	0.3623084	0.3427388	0.1773021	0.102888324	0.0362643	0.024616	0.0003689	0.0001596	0.0001234	2.888E-05	8.586E-06	2.911E-08
Cumulativ	56.136974	92.589406	96.086315	97.532097	98.483581	98.953193	99.315501	99.65824	99.835542	99.93843023	99.974695	99.999311	99.999679	99.999839	99.999963	99.999991	100	100

Standardized:

Variances																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Variance	5.5922409	4.7891836	1.2324537	1.0666955	0.9817471	0.7634697	0.696736	0.5972557	0.5385655	0.439214472	0.3967157	0.3418312	0.2102674	0.1915989	0.0903054	0.0357079	0.0216579	0.0143532
Variance I	31.068005	26.606576	6.8469653	5.9260861	5.4541508	4.2414981	3.8707556	3.3180872	2.9920306	2.440080398	2.2039761	1.8990624	1.1681525	1.0644385	0.5016969	0.1983773	0.1203217	0.0797402
Cumulativ	31.068005	57.674581	64.521546	70.447632	75.901783	80.143281	84.014036	87.332124	90.324154	92.76423451	94.968211	96.867273	98.035425	99.099864	99.601561	99.799938	99.92026	100

Problem 4.4 15 points (see 4.4_ToyotaCorolla.xlsx for data set)

a. The categorical variables are: model(kind of), fuel type, and color

b. When creating dummy variables for a categorical variables you are essentially creating a binary variable for each possible value (category) in the categorical variable. See part (d) for an example of this.

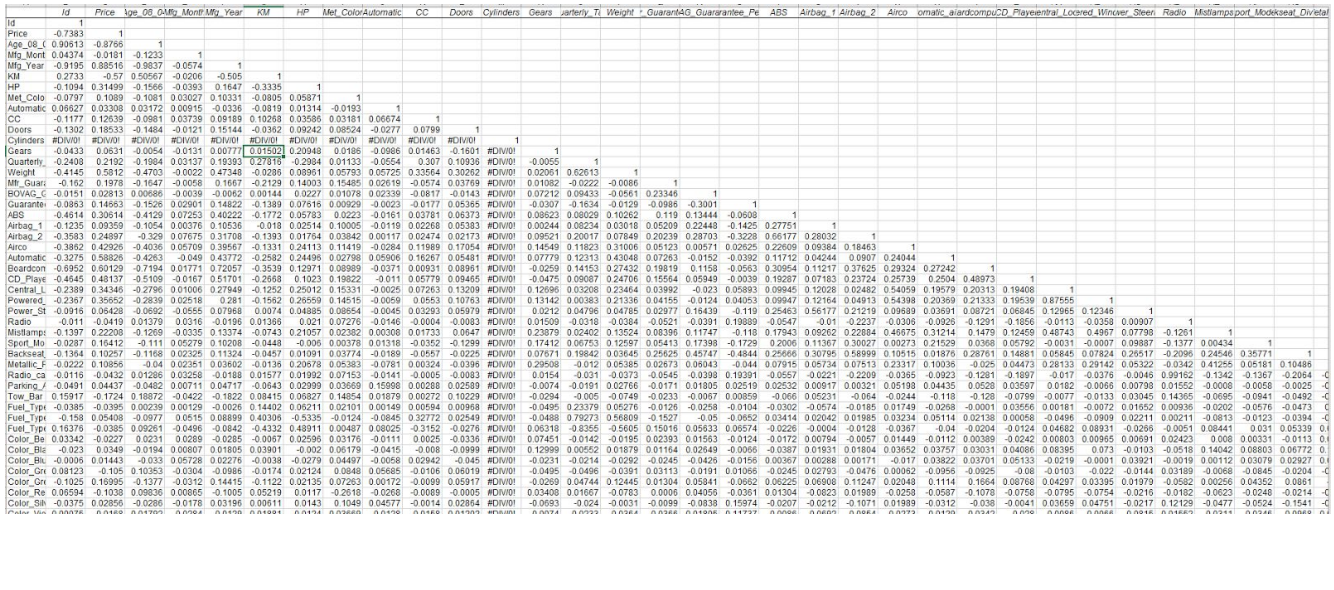
c. With a categorical variable with N categories, you can create N dummy variables, or N-1 dummy variables. for N-1 dummy variables, if all variables have a 0 value (not true), then that means the last category (which you did not create a dummy variable for) is true.

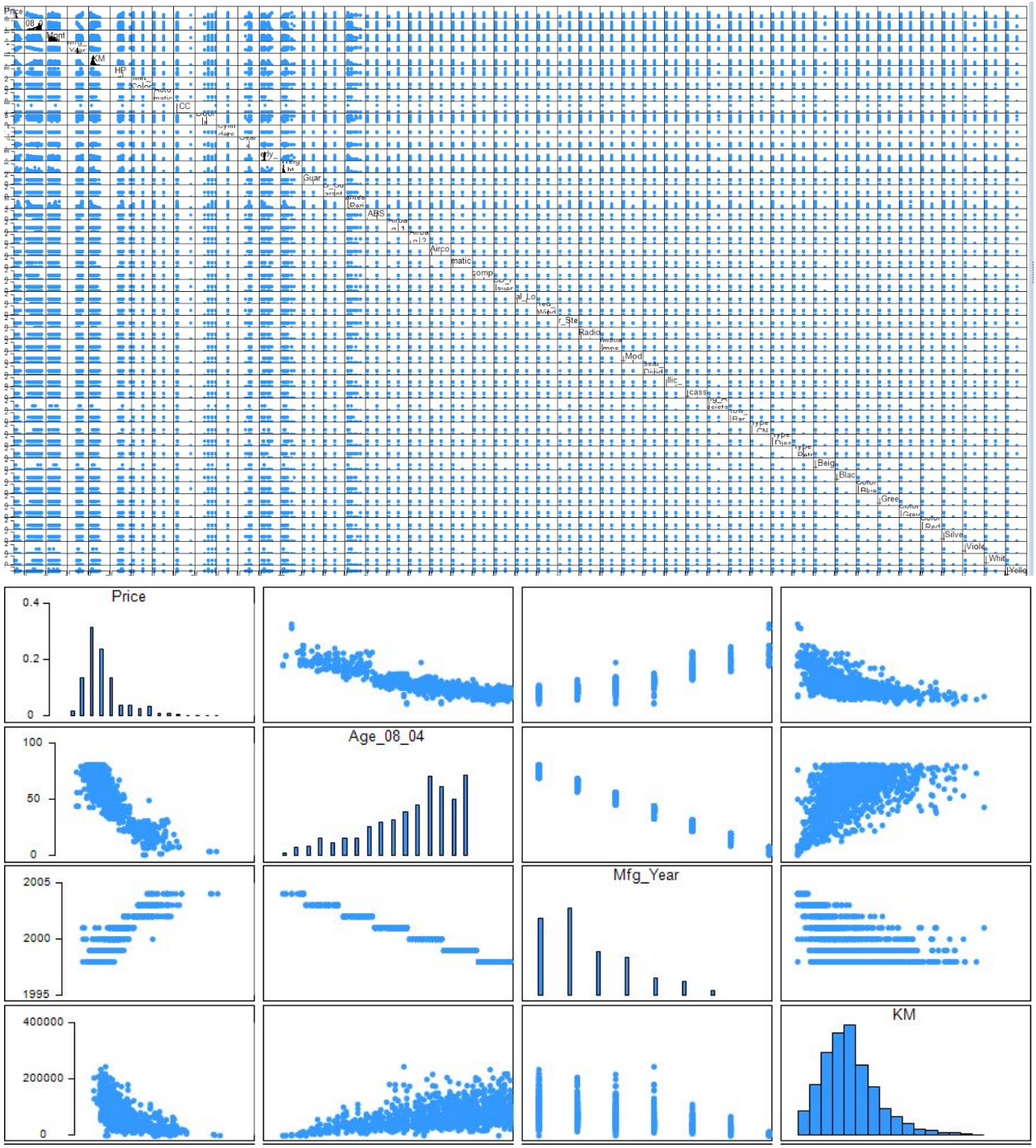
d. We find that there are 3 types of fuel in the Fuel Type column: Diesel, Petrol, and CNG. To prepare the data we would create a column for Diesel and Petrol, which would have binary values: 1 if true, 0 if false. If the value in the Diesel column is 1, then the fuel type the car used is Diesel; if the value is 0, it is not Diesel. The same goes for Petrol. If both Diesel and Petrol are 0, this means CNG is true.

AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX
_Assist	Tow_Bar	el_Type_Ch	l_Type_Die	l_Type_Pe	Color_Beige	Color_Black	Color_Blue	Color_Green	Color_Grey	Color_Red	Color_Silve	Color_Viole	Color_Whit	Color_Yellow
0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	1	0	0	0
0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0	1	0
0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	1	0	0	0
0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
0	0	0	0	1	0	0	0	0	0	1	0	0	0	0

e. Multiple pairs among the variables seem to be correlated, here are a few of the strongest ones (as there are a significant amount of relationships, we only listed a couple of them):

Price & Age
Price & Mfg_Yr
Mfg_Yr & Age
Price & KM




Price
Age_08_04
Mfg_Year
KM
0.4
0.2
0
100
50
0
2005
2000
1995
400000
200000
0