15

MATH 4044 – Statistics for Data Science

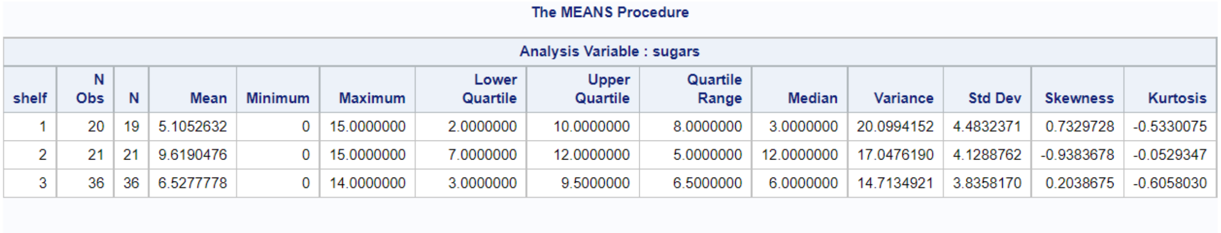
Student Information

Student ID Name: Wangjun SHEN

Student ID Number: 110248810

Answers

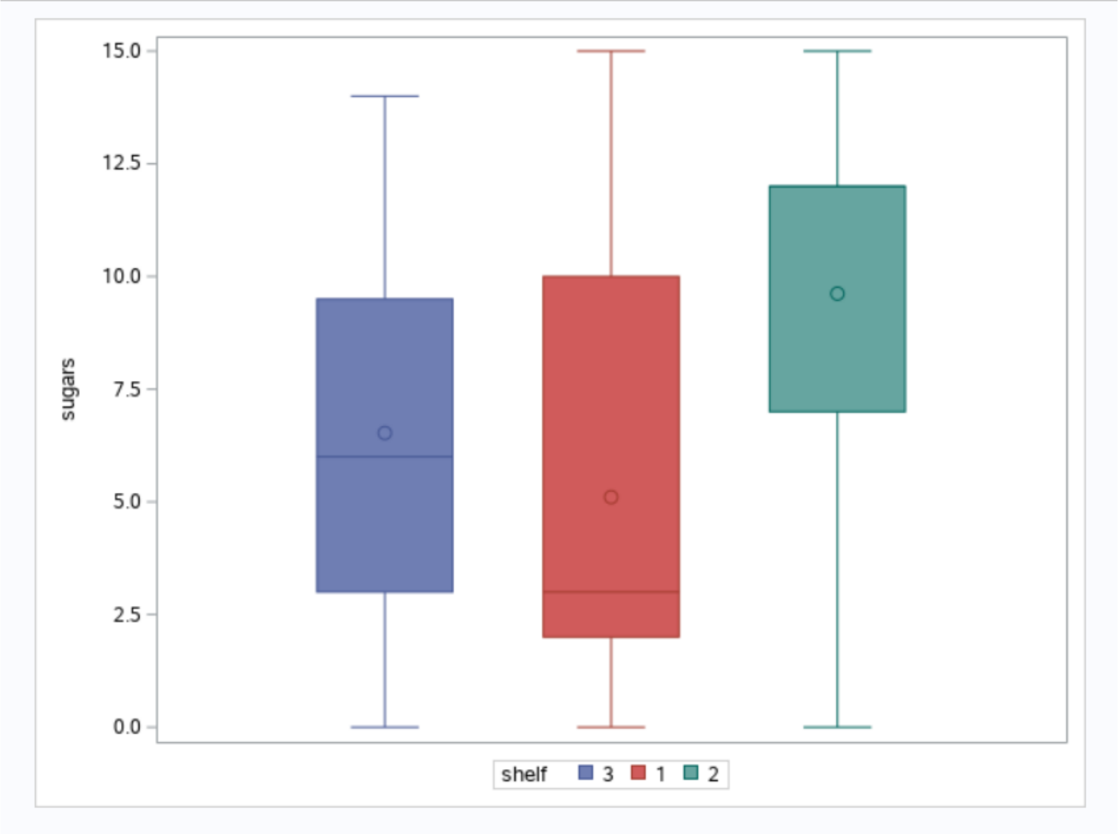
（a） Use SAS to study the distributions of sugar content by shelf location. More specifically, obtain measures of location, dispersion, skewness and kurtosis as well as boxplots and histograms, and use them to briefly describe, compare and contrast the distributions. Identify any outliers.

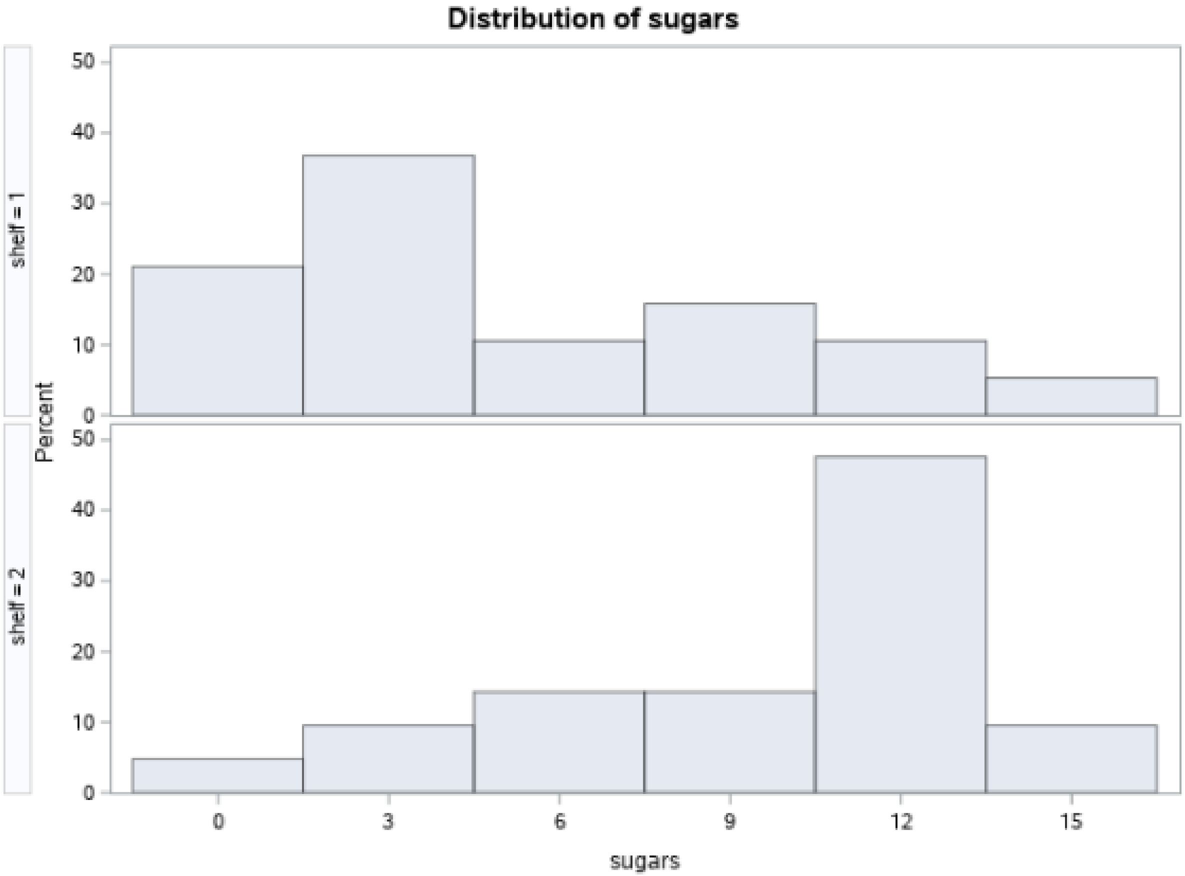


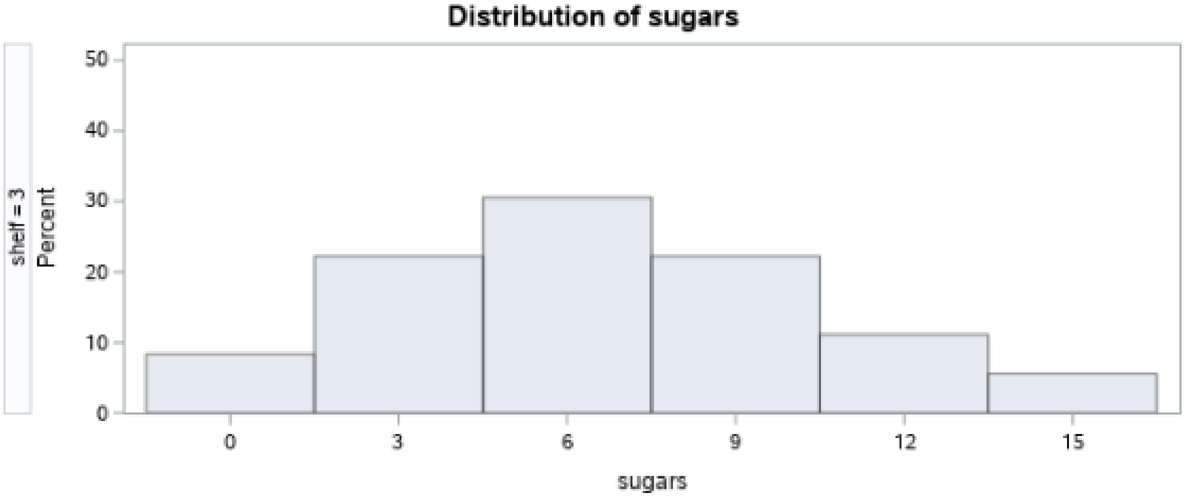
The MEANS Procedure

Analysis Variable : sugars

shelf	N Obs	N	Mean	Minimum	Maximum	Lower Quartile	Upper Quartile	Quartile Range	Median	Variance	Std Dev	Skewness	Kurtosis
1	20	19	5.1052632	0	15.0000000	2.0000000	10.0000000	8.0000000	3.0000000	20.0994152	4.4832371	0.7329728	-0.5330075
2	21	21	9.6190476	0	15.0000000	7.0000000	12.0000000	5.0000000	12.0000000	17.0476190	4.1288762	-0.9383678	-0.0529347
3	36	36	6.5277778	0	14.0000000	3.0000000	9.5000000	6.5000000	6.0000000	14.7134921	3.8358170	0.2038675	-0.6058030


15.0
12.5
10.0
7.5
5.0
2.5
0.0
sugars
shelf
🟦 3
🟥 1
🟩 2





1. The minimum and maximum values of the sugar content of shelf 1, 2, and 3 are not significantly different. The minimum value of their sugar content is 0, and the difference of their maximum value is only 1 while both shelf 1 and shelf 2 are 15 and only shelf 3 is 14.

2. The breakfast cereals' sugar content on shelf 2 has the highest mean value，which is nearly 88% higher than shelf 1's mean value and around 47% higher than shelf 3' mean value.

3. The breakfast cereals' sugar content on shelf 2 has the highest median value, which is exactly 4 times than shelf 1’s median value and 2 times than shelf 3’s median value.

4. From the value of standard deviation, the value of standard deviation of shelf 1 is the largest, which shows that the degree of dispersion of sugar content on this shelf is the largest among the three shelves.

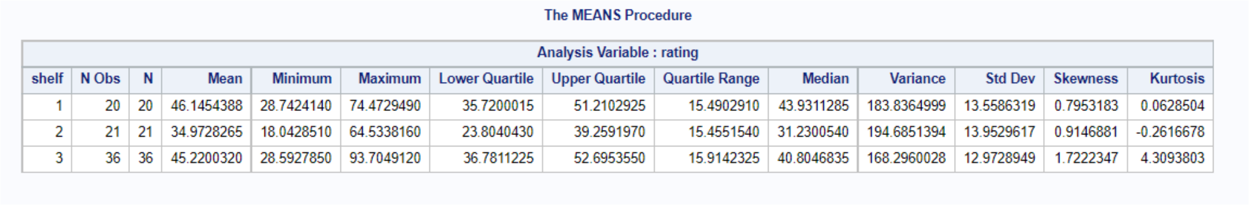
5. Shelf 1 is positively skewed, shelf 2 is negatively skewed, and shelf 3 is nearly positively skewed.

Is this confirmed by the boxplot and histograms?

6. From the picture, there should be no outliers.

Interpret kurtosis.

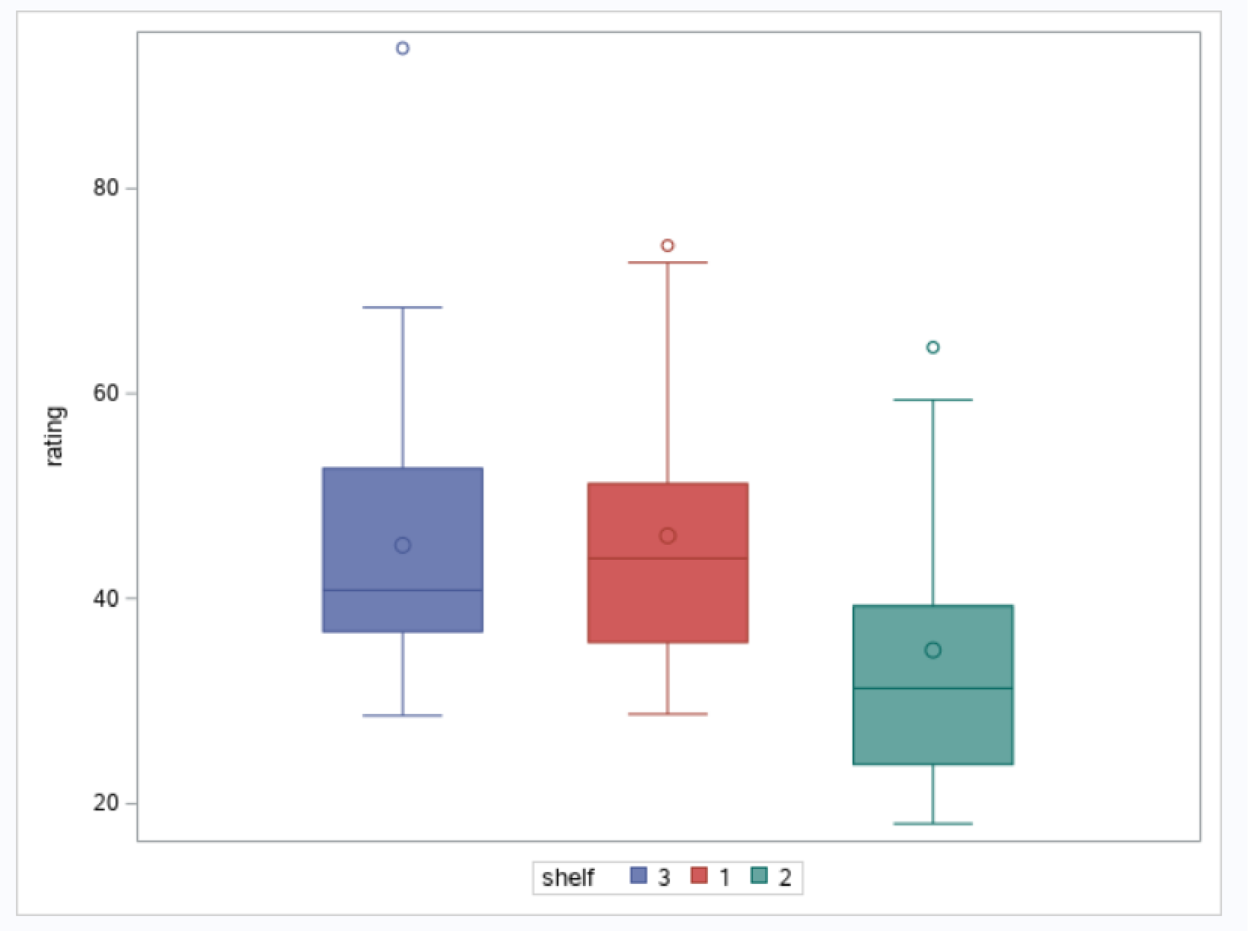
(b)Repeat part (a) for the distributions of health ratings by shelf location.

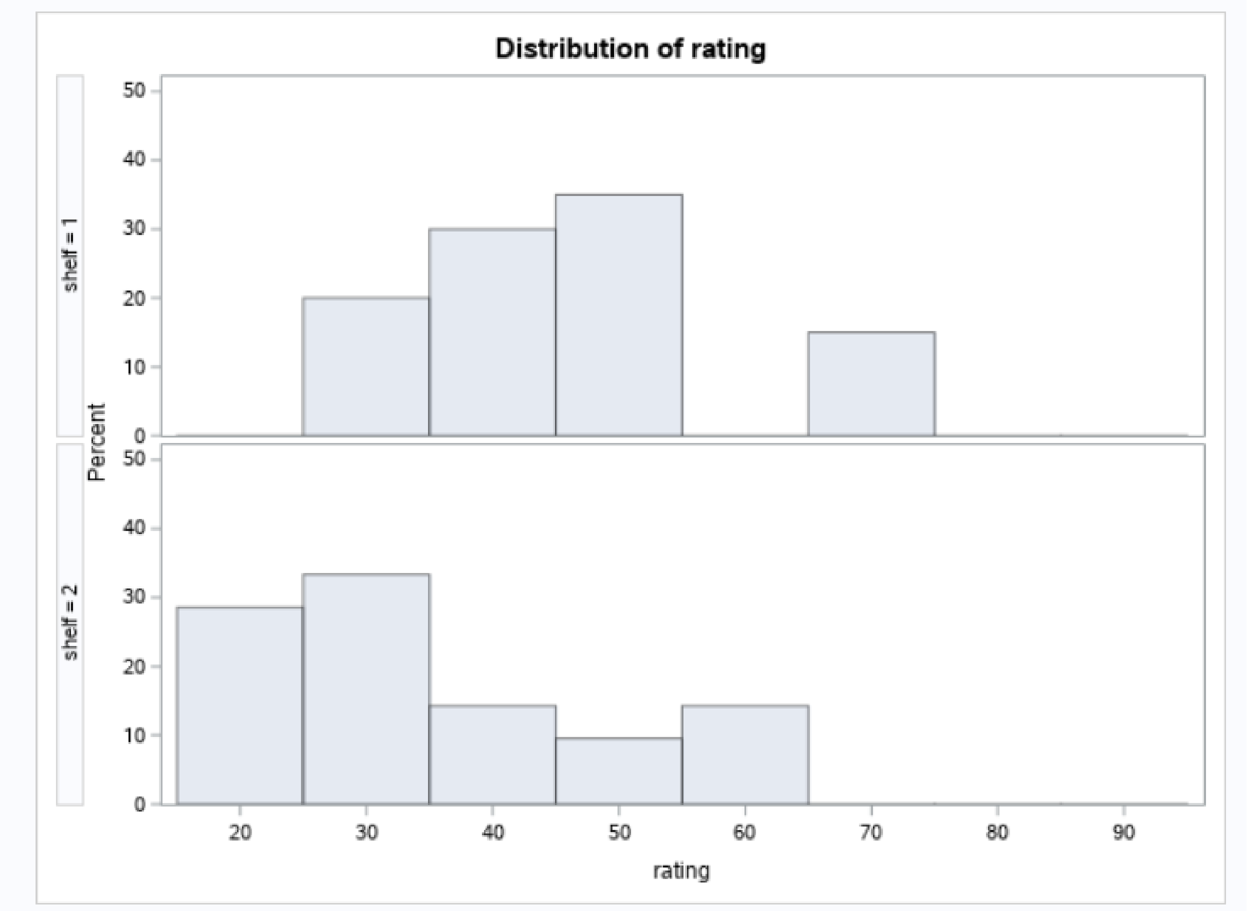


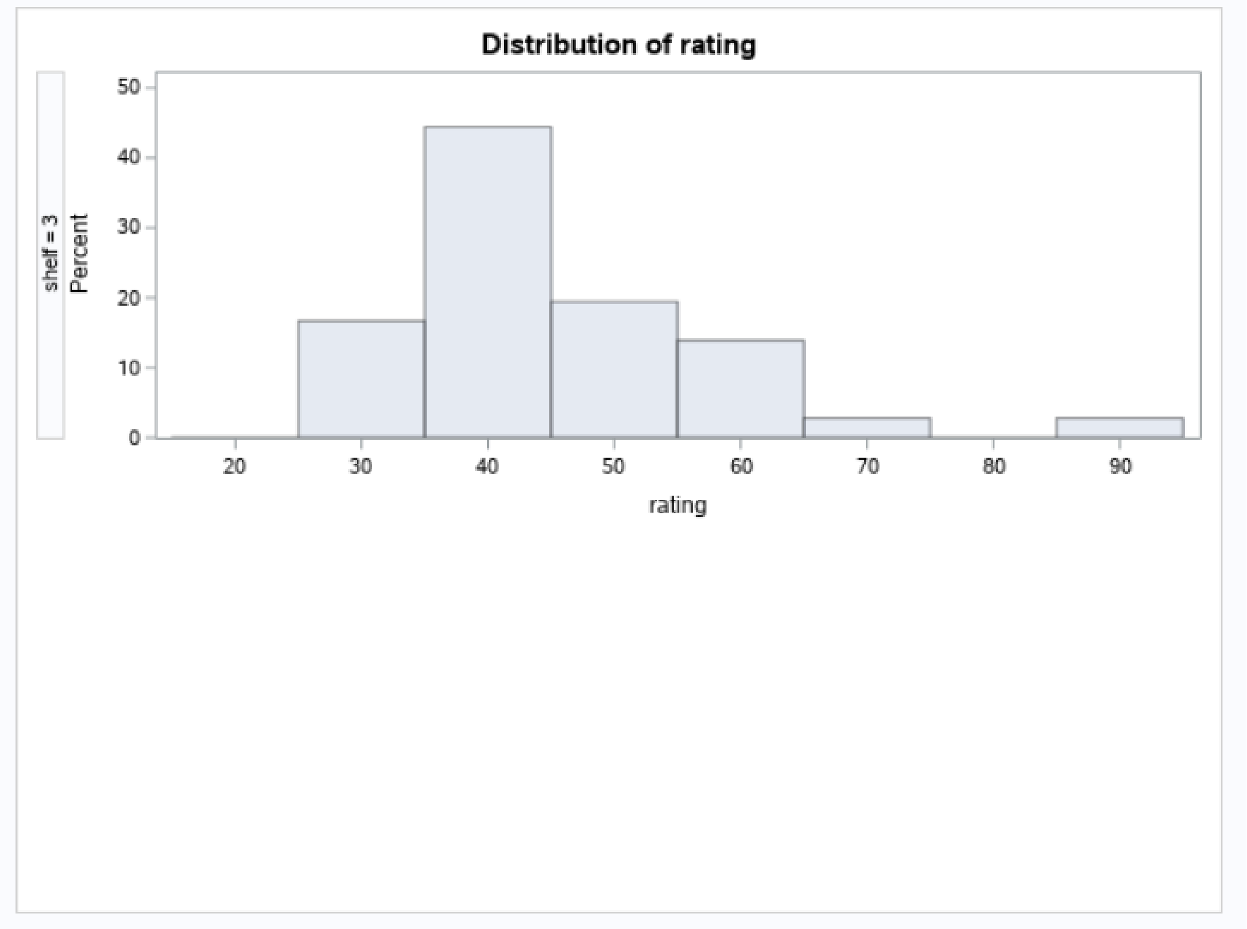
The MEANS Procedure

Analysis Variable : rating

shelf	N Obs	N	Mean	Minimum	Maximum	Lower Quartile	Upper Quartile	Quartile Range	Median	Variance	Std Dev	Skewness	Kurtosis
1	20	20	46.1454388	28.7424140	74.4729490	35.7200015	51.2102925	15.4902910	43.9311285	183.8364999	13.5586319	0.7953183	0.0628504
2	21	21	34.9728265	18.0428510	64.5338160	23.8040430	39.2591970	15.4551540	31.2300540	194.6851394	13.9529617	0.9146881	-0.2616678
3	36	36	45.2200320	28.5927850	93.7049120	36.7811225	52.6953550	15.9142325	40.8046835	168.2960028	12.9728949	1.7222347	4.3093803


80
60
40
20
rating
shelf
3
1
2


Distribution of rating
shelf = 1
shelf = 2
Percent
0
10
20
30
40
50
rating
20
30
40
50
60
70
80
90



1. The health rating of shelf 1 is very close to the health rating of shelf 2, but the mean value of health rating of ✓ shelf 2 is significantly lower than the other two. Likewise, Minimum value.
2. shelf 3d maximum value is significantly higher than the other two shelves
3. Shelf 1 and ✓ shelf 3d median values are very close, while shelf 2 has the lowest median value.
4. The standard deviations of the three shelves are relatively close, which indicates that the discrete lengths of the ✓ health ratings of the cereals on the three shelves are not much different.
5. All three shelves have outliers:
   a. shelf 1：74.4729490
   b. shelf 2：64.5338160
   c. shelf 3：93.7049120 ✓

HOw do you get this? Skewness values say otherwise?

6. Shelf 1 and self 3 are close to symmetrical, but there are outliers. ✓
7. Because there are outliers, IQR and Median are used. ✓

Interpret kurtosis.

(c) Based on your results from parts (a) and (b), what are your conclusions regarding sugar content and ratings of cereals, and their shelf location? One to two short paragraphs is sufficient.

Shelf 1 intersects the other two shelves and has a higher health rating because shelf 1 has a relatively lower sugar content.

Statistically speaking, the cereals of shelf 1 are healthier than the cereals of shelf 2 and shelf 3. ✓

SAS Code

```
libname mydata "~/MYDATA";

proc means data=mydata.cereals N mean min max q1 q3 qrange median var stddev skewness kurtosis;
  class shelf;
  var sugars;
run;


proc sgplot data=mydata.cereals;
  vbox sugars/group=shelf;
run;


proc univariate data=mydata.cereals;
  var sugars;
  class shelf;
  histogram;
  ods select histogram;
run;
```

```
* ==========================;

proc means data=mydata.cereals N mean min max q1 q3 qrange median var stddev skewness kurtosis;
  class shelf;
  var rating;
run;


proc sgplot data=mydata.cereals;
  vbox rating/group=shelf;
run;


proc univariate data=mydata.cereals;
  var rating;
  class shelf;
  histogram;
  ods select histogram;
run;
```