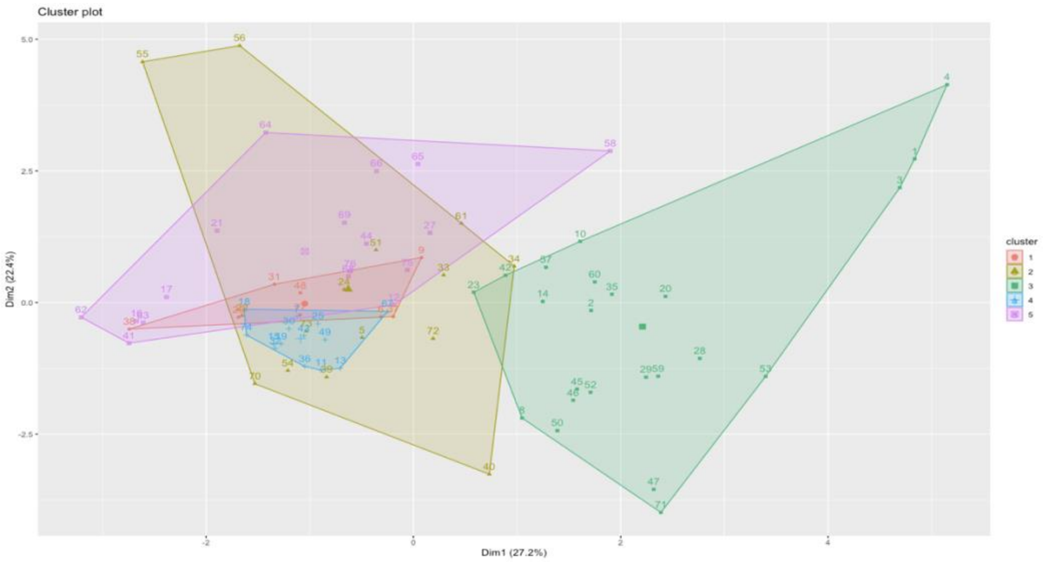
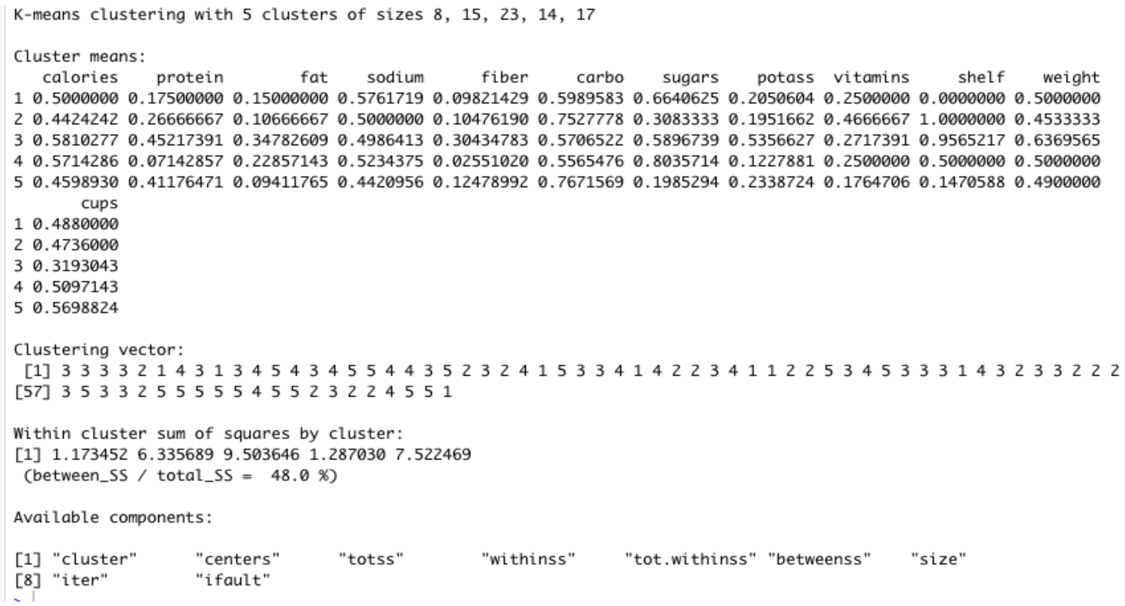
1. **Using all of the variables, except name and rating, run the k-means algorithm with k=5 to identify clusters within the data.**

Cluster plot



K-means metrics output

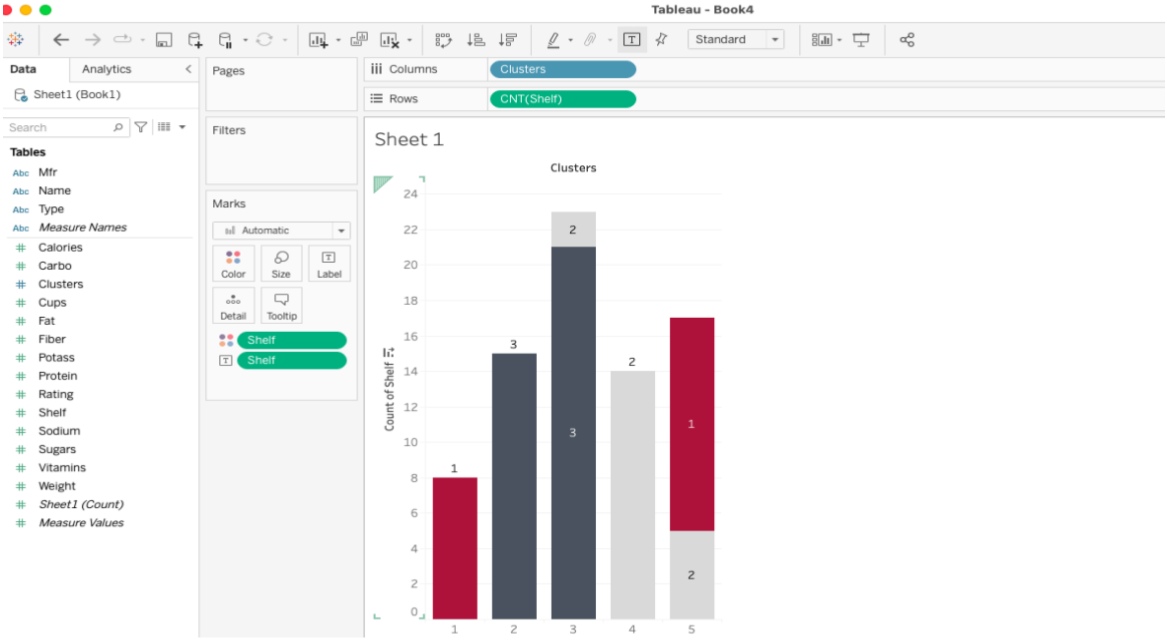


```
K-means clustering with 5 clusters of sizes 8, 15, 23, 14, 17

Cluster means:
   calories    protein        fat    sodium      fiber     carbo    sugars    potass  vitamins     shelf    weight
1 0.5000000 0.17500000 0.15000000 0.5761719 0.09821429 0.5989583 0.6640625 0.2050604 0.2500000 0.0000000 0.5000000
2 0.4424242 0.26666667 0.10666667 0.5000000 0.10476190 0.7527778 0.3083333 0.1951662 0.4666667 1.0000000 0.4533333
3 0.5810277 0.45217391 0.34782609 0.4986413 0.30434783 0.5706522 0.5896739 0.5356627 0.2717391 0.9565217 0.6369565
4 0.5714286 0.07142857 0.22857143 0.5234375 0.02551020 0.5565476 0.8035714 0.1227881 0.2500000 0.5000000 0.5000000
5 0.4598930 0.41176471 0.09411765 0.4420956 0.12478992 0.7671569 0.1985294 0.2338724 0.1764706 0.1470588 0.4900000
       cups
1 0.4880000
2 0.4736000
3 0.3193043
4 0.5097143
5 0.5698824

Clustering vector:
 [1] 3 3 3 3 2 1 4 3 1 3 4 5 4 3 4 5 5 4 4 3 5 2 3 2 4 1 5 3 3 4 1 4 2 2 3 4 1 1 2 2 5 3 4 5 3 3 3 1 4 3 2 3 3 2 2 2
[57] 3 5 3 3 2 5 5 5 5 5 4 5 5 2 3 2 2 4 5 5 1

Within cluster sum of squares by cluster:
[1] 1.173452 6.335689 9.503646 1.287030 7.522469
 (between_SS / total_SS =  48.0 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"
```

2. **Develop clustering profiles that clearly describe the characteristics of the cereals within the cluster.**

Based on the above result, we can see that – cluster defining metrics are majorly **shelf** and **sugar** since their cluster means are varying highly between clusters. The next three important metrics defining the cluster are - sodium, potassium, fiber. Others are not contributing much in defining the cluster.

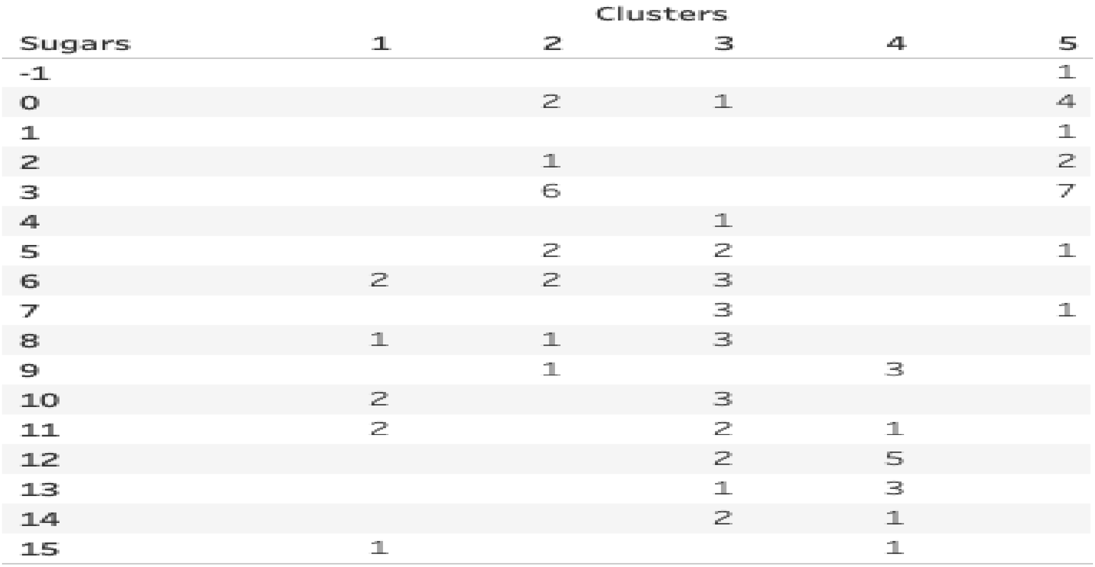
Below are the findings based on the analysis –



Columns Clusters

Rows CNT(Shelf)

Below table represents the number of cereals belonging to different clusters and different sugar levels.



			Clusters		
Sugars	1	2	3	4	5
-1					1
0		2	1		4
1					1
2		1			2
3		6			7
4			1		
5		2	2		1
6	2	2	3		
7			3		1
8	1	1	3		
9		1		3	
10	2		3		
11	2		2	1	
12			2	5	
13			1	3	
14			2	1	
15	1			1	

From the graphs, Based on cluster –

1. Cluster 1 has only shelf 1
2. Cluster 2 has only shelf 3

3. Cluster 4 has only shelf 2
4. Cluster 3 has shelf 2 and 3
5. Cluster 5 has shelf 1 and 2

Also, Cluster 1,2,3,4 contains Cold type and only cluster 5 contains very few hot.

Based on shelf and sugar –

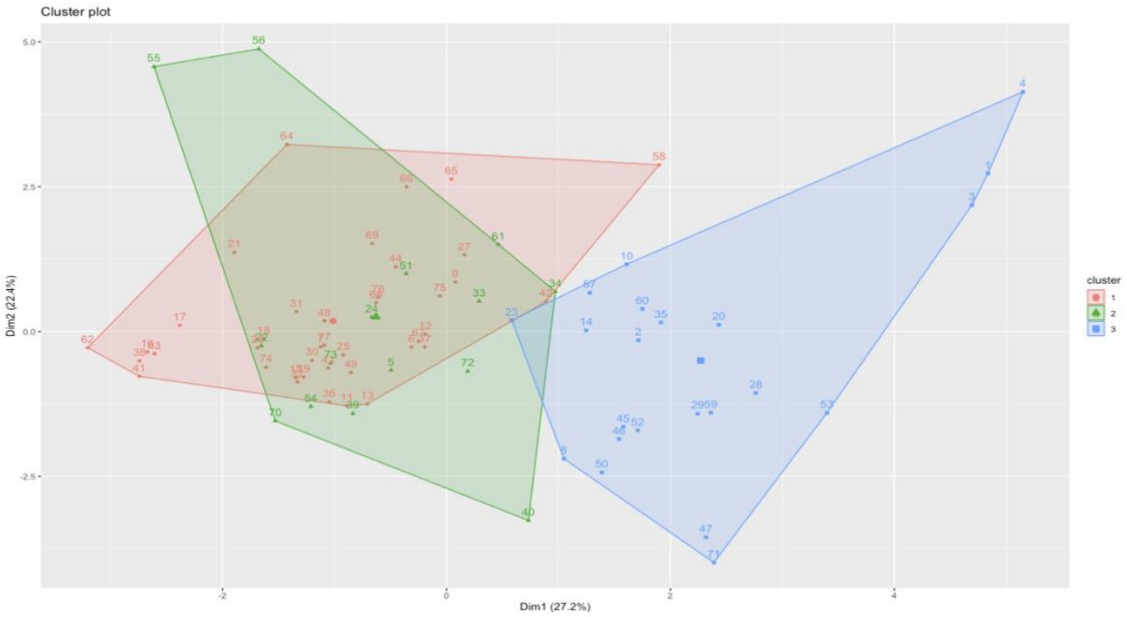
Shelf 1 which has sugar >5 belong to cluster 1
Shelf 1 which has sugar <5 belong to cluster 5

Shelf 2 which has sugar in range from 0 to 7 belong to cluster 5
Shelf 2 which has sugar in range from 6 to 12 belong to cluster 3
Shelf 2 which has sugar in range from 9 to 15 belong to cluster 4

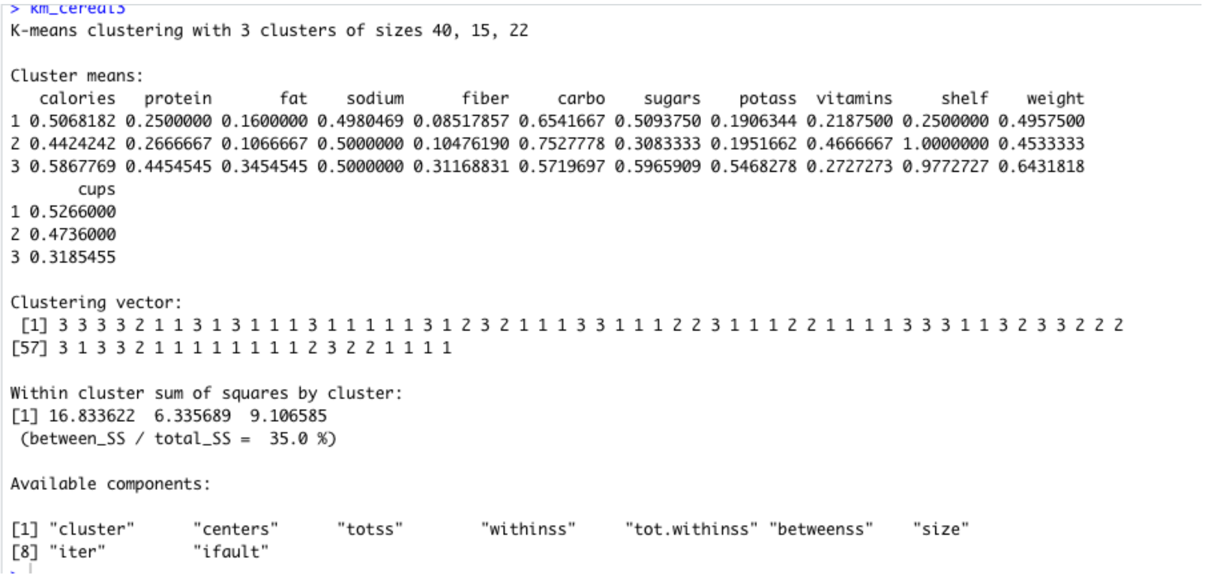
Shelf 3 belongs to cluster 2 and 3
Since, sugar cannot independently predict the cluster, we look at the next best predictors which are sodium, potassium, fibre. When sugar is same/common across both clusters, then Lesser the sodium, potassium, fibre – it belong to cluster 2. Otherwise it belongs to cluster 3.

3. Rerun the k-means algorithm with k=3.

Cluster plot



K-means metrics output



```
> km_cereal3
K-means clustering with 3 clusters of sizes 40, 15, 22

Cluster means:
   calories   protein       fat    sodium      fiber     carbo    sugars    potass  vitamins     shelf    weight
1 0.5068182 0.2500000 0.1600000 0.4980469 0.08517857 0.6541667 0.5093750 0.1906344 0.2187500 0.2500000 0.4957500
2 0.4424242 0.2666667 0.1066667 0.5000000 0.10476190 0.7527778 0.3083333 0.1951662 0.4666667 1.0000000 0.4533333
3 0.5867769 0.4454545 0.3454545 0.5000000 0.31168831 0.5719697 0.5965909 0.5468278 0.2727273 0.9772727 0.6431818
       cups
1 0.5266000
2 0.4736000
3 0.3185455

Clustering vector:
 [1] 3 3 3 3 2 1 1 3 1 3 1 1 1 3 1 1 1 1 1 3 1 2 3 2 1 1 1 3 3 1 1 1 2 2 3 1 1 1 2 2 1 1 1 1 3 3 3 1 1 3 2 3 3 2 2 2
[57] 3 1 3 3 2 1 1 1 1 1 1 1 1 2 3 2 2 1 1 1 1

Within cluster sum of squares by cluster:
[1] 16.833622  6.335689  9.106585
 (between_SS / total_SS =  35.0 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"
```

4. Which clustering solution do you prefer, and why?

I would go with the cluster solution where k=5 since the between_SS/total_SS value is greater than k=3.

1. between_SS/total_SS= 35% (k=3)
2. between_SS/total_SS= 48% (k=5)

For good cluster characteristics, between cluster variance should be high and within cluster variance should be low. Looking at the k-means clustering summary output where k=5, we can see that it is far more easier to separate and understand the clusters with respect to shelf, sugar metrics.

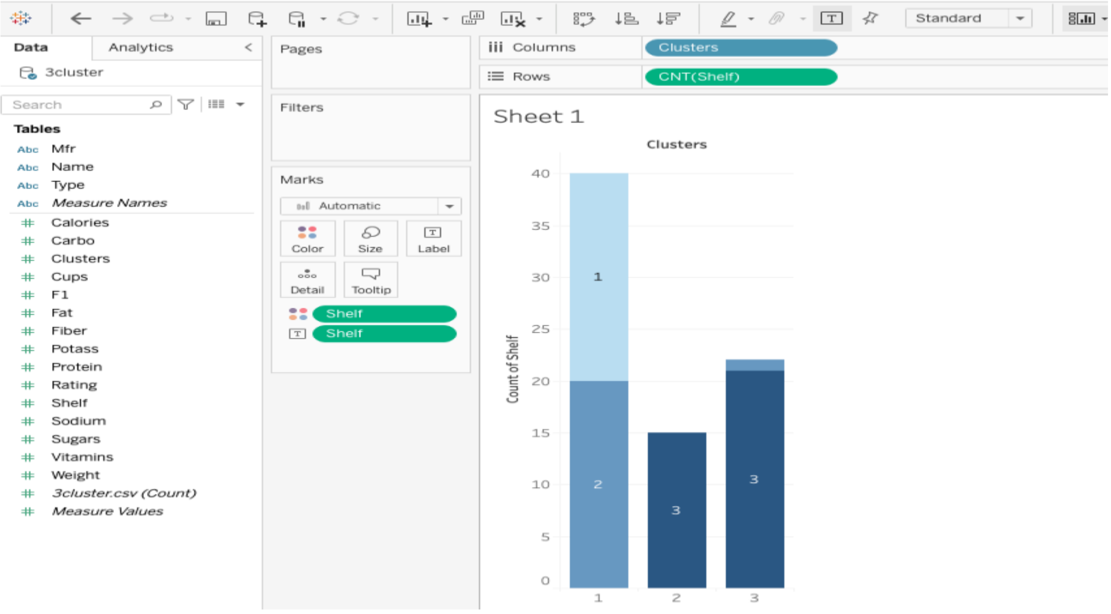
5. Develop clustering profiles that clearly describe the characteristics of the cereals within the cluster.

Clustering profiles for k=3:

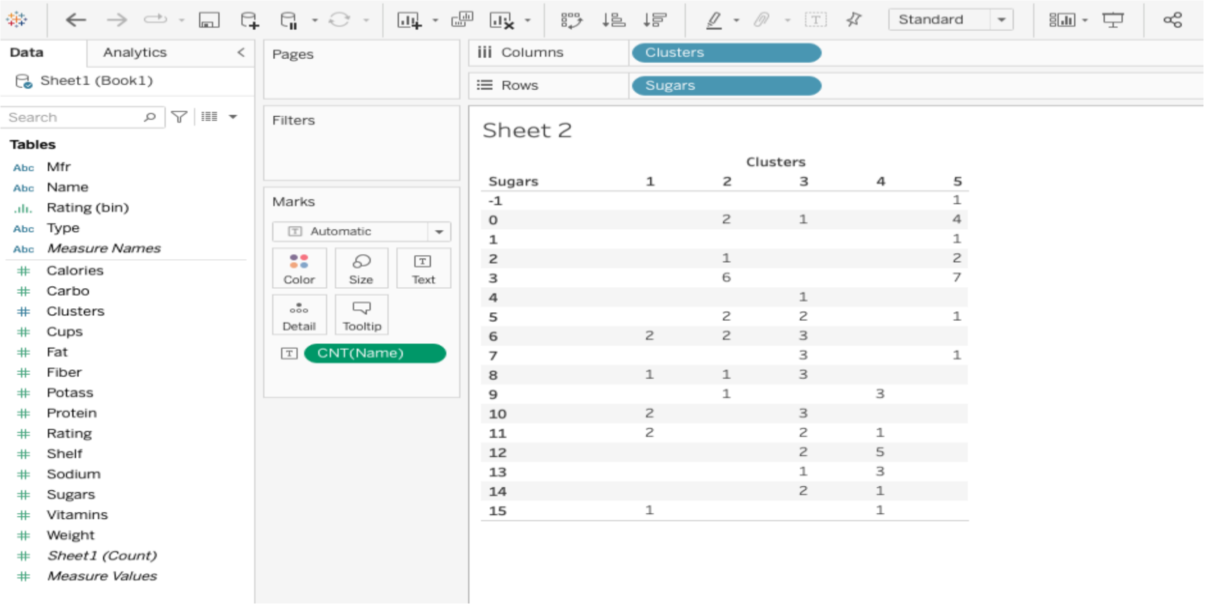
Based on the output of k-means attached in question 3, we can infer that **shelf**, **sugar**, potassium, fiber are contributing mainly in defining the cluster.

1. Cluster 1 has only shelf 1 & 2 (50% shelf 1 and 50% shelf 2)
2. Cluster 2 has only shelf 3 (100% shelf 3)
3. Cluster 3 has shelf 2 and 3 (4% shelf 2 and 96% shelf 3)

Cluster with respect to shelf:



Below table represents the number of cereals belonging to different clusters and different sugar levels.



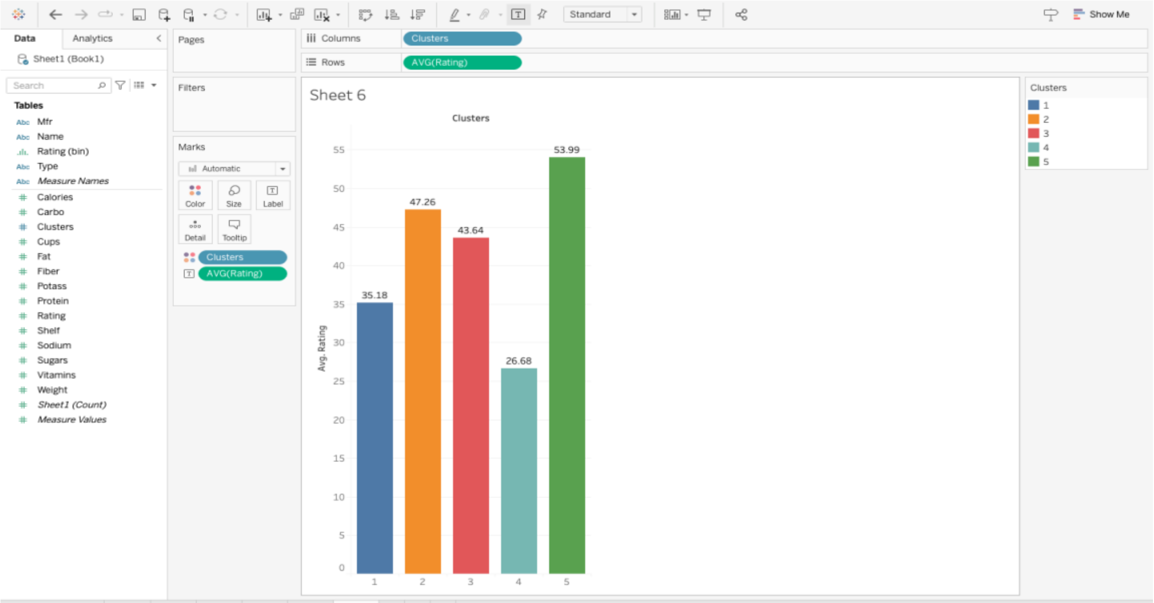
Sheet 2

	Clusters				
Sugars	1	2	3	4	5
-1					1
0		2	1		4
1					1
2		1			2
3		6			7
4			1		
5		2	2		1
6	2	2	3		
7			3		1
8	1	1	3		
9		1		3	
10	2		3		
11	2		2	1	
12			2	5	
13			1	3	
14			2	1	
15	1			1	

6. **Use cluster membership to predict rating. One way to do this would be to construct a histogram of rating based on cluster membership alone. Describe how the relationship**

you uncovered makes sense, based on your earlier profiles.

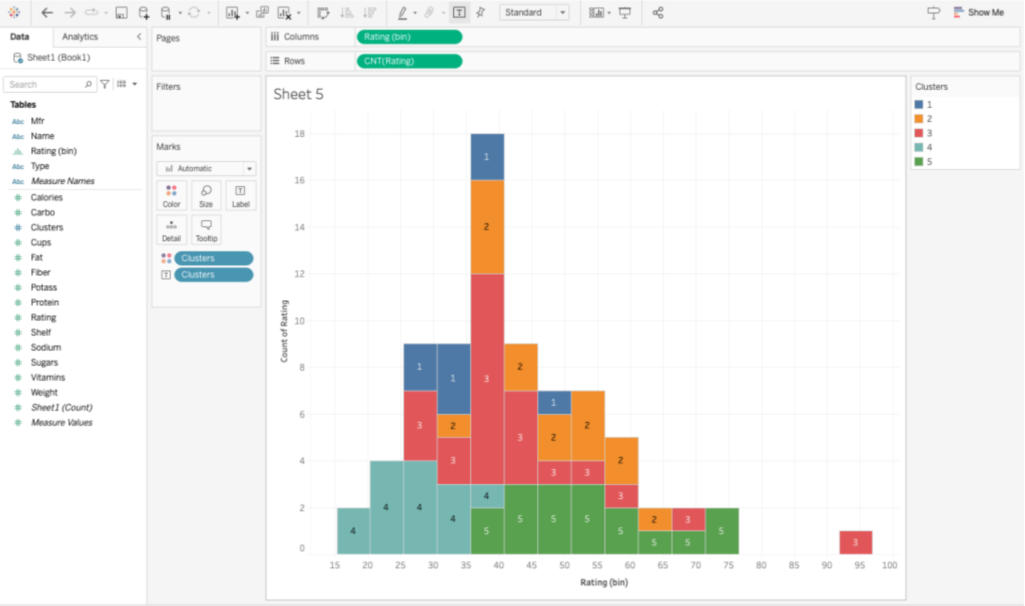
Below shows the average rating of cereals for different clusters:



Average cluster ratings for all the 5 clusters –

Cereals with average rating of 35.18 belongs to cluster 1
Cereals with average rating of 47.26 belongs to cluster 2
Cereals with average rating of 43.64 belongs to cluster 3
Cereals with average rating of 26.68 belongs to cluster 4
Cereals with average rating of 53.99 belongs to cluster 5

Below is the histogram of cereals with respect to ratings –


Data
Analytics
Sheet1 (Book1)
Search
Tables
Mfr
Name
Rating (bin)
Type
Measure Names
Calories
Carbo
Clusters
Cups
Fat
Fiber
Potass
Protein
Rating
Shelf
Sodium
Sugars
Vitamins
Weight
Sheet1 (Count)
Measure Values
Pages
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Clusters
Clusters
Columns
Rating (bin)
Rows
CNT(Rating)
Standard
Show Me
Sheet 5
Count of Rating
Rating (bin)
Clusters