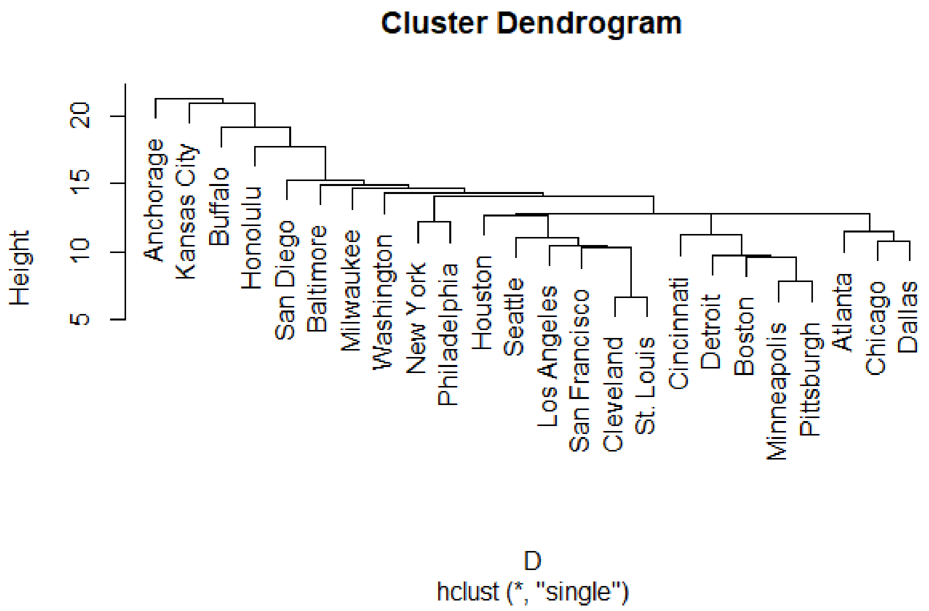
Vandana Agrawal

Cluster Analysis- Food Data Set

Cluster Analysis is the data exploration method used for grouping the similar data point into groups.There are various method to identify how to group the data point in a group like Single link Clustering, K mean Clustering.

Single Link Clustering : Single Link Clustering also known as minimum distance clustering or nearest neighbor where distance between the cluster is defined by the minimum distance between the data point of the two cluster.

Single Link Clustering on Unstandardized data



```
Cleveland      38.5      107.7 142.70     50.3     83.2
St. Louis      36.9      109.8 140.00     46.7     79.0
```

From the Dendrogram we can conclude that Cleveland and St. louis has the shorted distance meaning that both cities are almost similar in terms of their food prices. We can verify that by checking the food prices for example Bread in Cleveland i s of 38.5 cents and 36.9 cent in St.Louis. So, if we calculate the Euclidian distance = sqrt((38.5-36.9)^2) +((107.7-109.8)^2)+ ((142.7-140.0)^2)+ ((50.3-46.7)^2)+ ((83.2-79.0)^2)) =6.69

```
Minneapolis    32.5      116.7 135.10     48.0     89.1
Pittsburgh     36.9      115.4 138.90     43.8     91.9
```

Distance between above two cities is 7.808329. So, Minneapolis and Pittsburgh are similar. From the distance matrix Anchorage is the city which has high distance value with each of the cities almost all the value of the distance is in 40’s and 50’s And Also from the dendrogram it appears that Anchorage is the outlier.

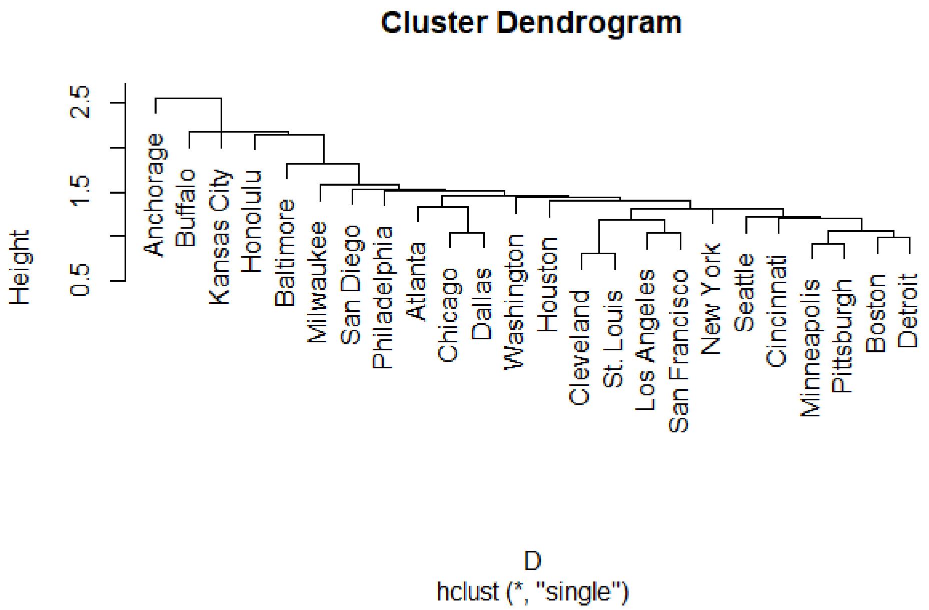
Vandana Agrawal

The mean price of individual food item differs a lot like mean price of bread is 38.44 where Mean price of Butter is 144.2. Euclidean distance will be influenced by butter price and Hence we should do standardization of the data in this case. The total global error for unstandardized data is

```
> tss_single
[1] 348551.5
```

Single Link Clustering Standardized data

Three Cluster:

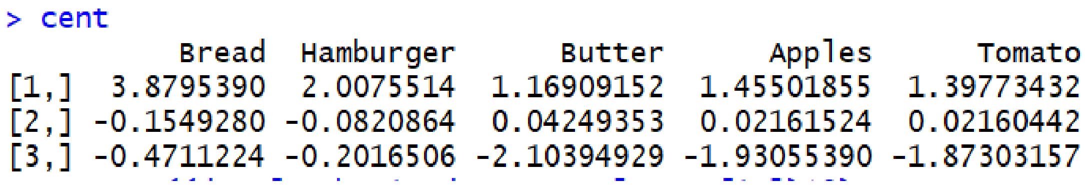


After Standardizing the data, we again calculate the Euclidian distance and plot dendrogram. From the dendrogram we can see that Cleveland and St Louis are the most similar cities we can replace one with other. There are other cities that have the shortest distance between them, for example: Minneapolis and Pittsburgh have the distance .92, Boston- Detroit is .98. Los Angeles- San Francisco is 1.04

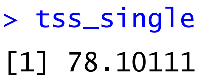
When we choose three cluster for single cluster analysis, we found that Anchorage is in cluster 1, Buffalo is in cluster 3 and rest all cities are in Cluster 2. We can say that the cluster 1 represents the city which has the highest food price of all the food items. Cluster 3 contain city having lowest price of all the food items. And Cluster 2 represents cities having low price of bread and Hamburger and high price of apple, tomato and butter

```
> memb
    Anchorage       Atlanta     Baltimore        Boston       Buffalo       Chicago    Cincinnati     Cleveland 
            1             2             2             2             3             2             2             2 
       Dallas       Detroit      Honolulu       Houston   Kansas City   Los Angeles     Milwaukee   Minneapolis 
            2             2             2             2             2             2             2             2 
     New York  Philadelphia    Pittsburgh     St. Louis     San Diego San Francisco       Seattle    Washington 
            2             2             2             2             2             2             2             2 
```

Vandana Agrawal



```
> cent
          Bread  Hamburger      Butter      Apples      Tomato
[1,]  3.8795390  2.0075514  1.16909152  1.45501855  1.39773432
[2,] -0.1549280 -0.0820864  0.04249353  0.02161524  0.02160442
[3,] -0.4711224 -0.2016506 -2.10394929 -1.93055390 -1.87303157
```

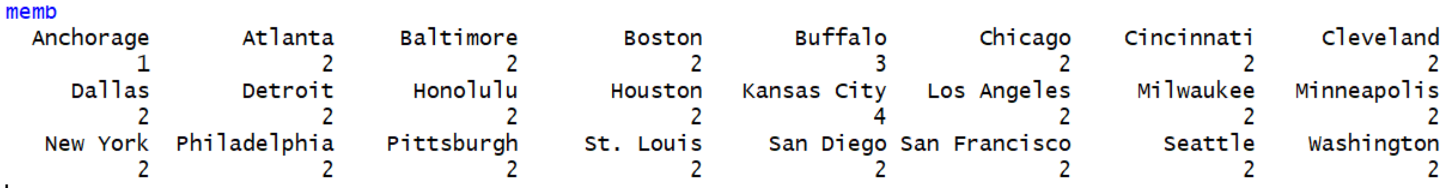


```
> tss_single
[1] 78.10111
```

Global Error for three cluster is 78.1. Global error is reduced by huge difference when we compare it with unstandardized data single link clustering with three cluster results.

Four Cluster:

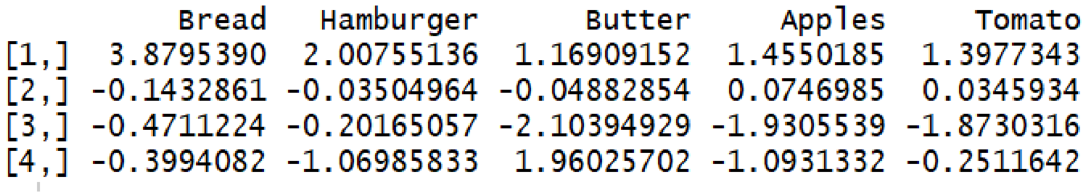
Now we are considered four cluster for single cluster analysis. We found that Anchorage is in Cluster 1. Buffalo is in Cluster 3, Kansas City is cluster 4 and rest all cities lies in cluster 2



```
memb
   Anchorage       Atlanta     Baltimore        Boston       Buffalo       Chicago    Cincinnati     Cleveland 
           1             2             2             2             3             2             2             2 
      Dallas       Detroit      Honolulu       Houston   Kansas City   Los Angeles     Milwaukee   Minneapolis 
           2             2             2             2             4             2             2             2 
    New York  Philadelphia    Pittsburgh     St. Louis     San Diego San Francisco       Seattle    Washington 
           2             2             2             2             2             2             2             2 
```

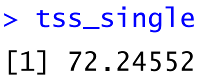
Identifying the centroid of each cluster. We see that the cluster one represents the city having highest food price.

Cluster 2 has cities having lower Bread, Hamburger and butter price and high Apple and tomato price. Cluster 3 has city with low food price. And Cluster 4 has city having high butter price and rest all food items are cheap.



```
          Bread   Hamburger      Butter     Apples     Tomato
[1,]  3.8795390  2.00755136  1.16909152  1.4550185  1.3977343
[2,] -0.1432861 -0.03504964 -0.04882854  0.0746985  0.0345934
[3,] -0.4711224 -0.20165057 -2.10394929 -1.9305539 -1.8730316
[4,] -0.3994082 -1.06985833  1.96025702 -1.0931332 -0.2511642
```

The global error for four cluster has improved and Total global error is now 72.2

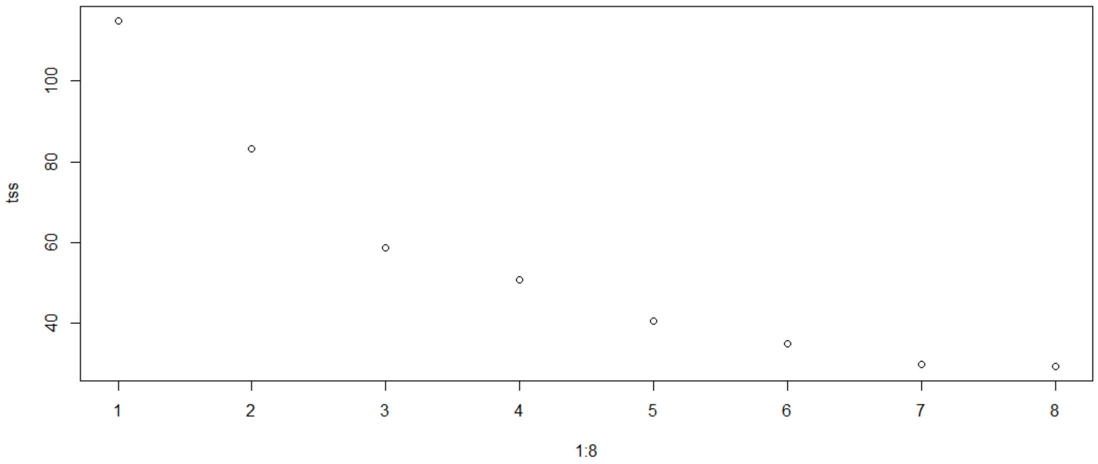


```
> tss_single
[1] 72.24552
```

Comparing Cluster 3 and cluster 4 results there is not much difference in terms of total global error there is approx. 6% decrease in error in Cluster 4

K- mean Cluster Analysis: Before we start the K mean Cluster analysis, we need to decide on how many clusters we want. In order to decide so we can plot the total sum of square for each cluster and see for which cluster we are getting tss with lowest value.

Vandana Agrawal



In the above graph we see that after 4 the curve is getting steady and there is not much difference in tss when we choose 5 or more number of clusters for analysis.

Three cluster in K mean Cluster Analysis.

Applying K-Mean Clustering to standardized data. Now Anchorage and Honolulu are in same cluster that is cluster 1. Buffalo, San Diego, Seattle, Houston and Sab Francisco are in the same cluster and rest all cities are in Cluster 2

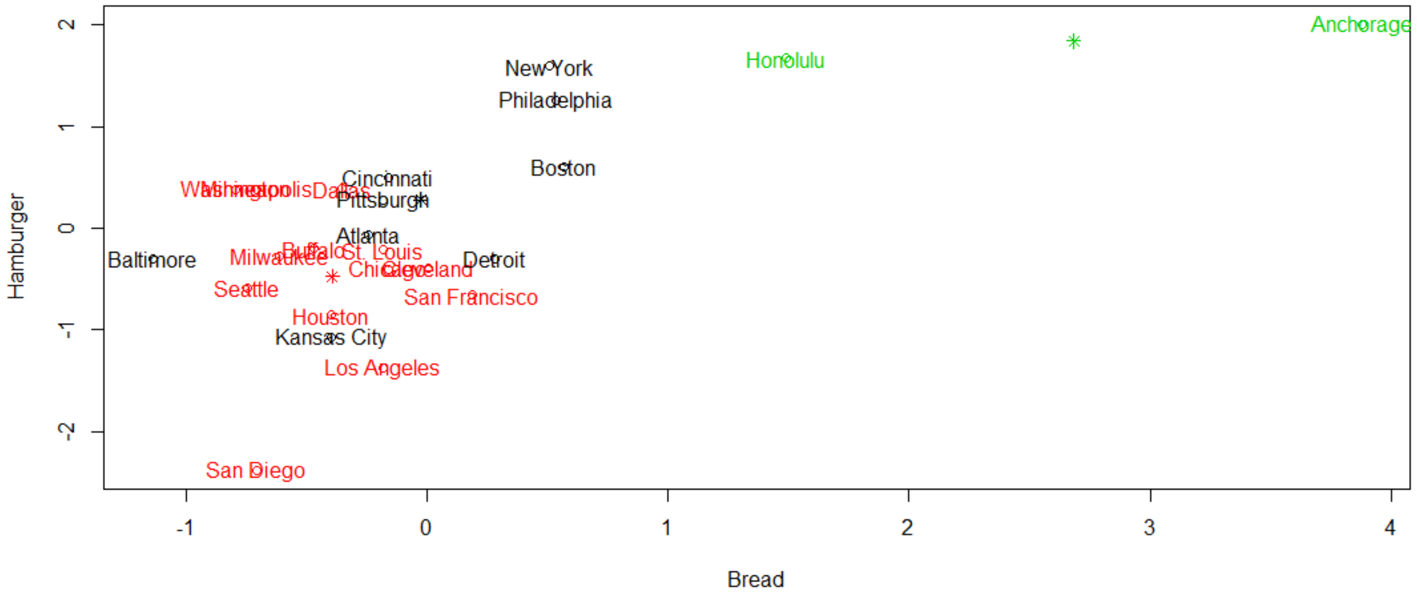
```
Clustering vector:
    Anchorage       Atlanta     Baltimore        Boston       Buffalo       Chicago    Cincinnati     Cleveland 
            1             2             2             2             3             2             2             3 
       Dallas       Detroit      Honolulu       Houston   Kansas City   Los Angeles     Milwaukee   Minneapolis 
            2             2             1             3             2             3             3             3 
     New York  Philadelphia    Pittsburgh     St. Louis     San Diego San Francisco       Seattle    Washington 
            2             2             2             3             3             3             3             2 
```

```
Cluster means:
       Bread   Hamburger     Butter       Apples      Tomato
1  2.6843023  1.8399271  1.1365779  1.52081589  0.9787519
2 -0.1284879  0.2396167  0.3870479 -0.21982295  0.5811691
3 -0.3826749 -0.6555255 -0.6917731 -0.04037564 -0.8931534
```

Cluster means gives us how these cities are grouped in a cluster, and we see that Cluster one has cities having high food price of all the food items Though we have same group in Single linage clustering, but Honolulu was not included that time. Cluster 2 has cities having low bread and Apple price. Cluster three has all these cities having with low food prices. K mean Clustering has more clearly sperate out the cities in clusters. In single linkage cities who are outliers were kept in one cluster. In k mean makes more sense it creates the clusters.

Total Sum of square with in is 58.8%. When we compare the global error of K mean clustering with single linkage the global error is reduced by 20%. We can say that with K mean we are get better quality clusters.

Vandana Agrawal



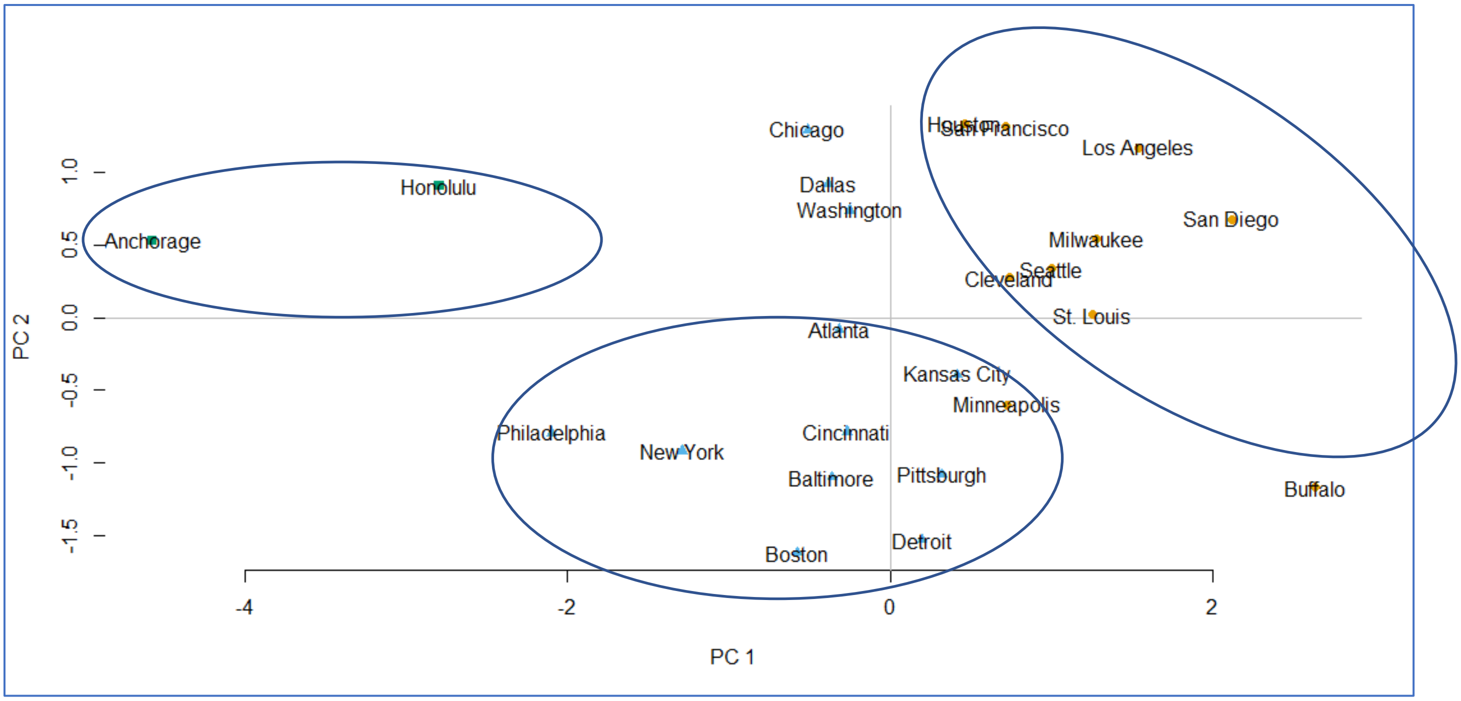
From the above graph we can deduce that Anchorage and Honolulu are in the cluster having High price of Bread and Hamburger. New York and Philadelphia re in the Cluster having High Hamburger price and Moderate Bread Price. Houston Seattle San Francisco are in the cluster where Hamburger and Bread price are low. San Diego looks like the outlier in the cluster having very low Bread price and Hamburger price compare to other cities in the same cluster.

Doing principal Component analysis to understand clustering

```
                PC1         PC2        PC3         PC4         PC5
Bread     -0.5099267  0.05649609 -0.4017162 -0.53197875 -0.54074549
Hamburger -0.5200718 -0.27761601 -0.4074371  0.07148075  0.69378685
Butter    -0.3973106  0.09940133  0.7684773 -0.43041438  0.23759156
Apples    -0.2909422  0.87675286 -0.0713631  0.37257819  0.05243928
Tomato    -0.4764421 -0.37571444  0.2774329  0.62275040 -0.40872301
```

So Principal Component 1 value will be more if Bread and Hamburger price less. And PC2 Price increase with Increase in price of Apple

Vandana Agrawal



From Above plot we can conclude that Anchorage and Honolulu plotted in low value of PCA1 and high value of PCA2 Meaning that these cities have high price of Bread Hamburger and Apple. Another Cluster we have with cities Atlanta, New York, Boston is the cluster with Low price of Apple (PCA2) and low to Moderate price of Hamburger and bread. Cluster with Los Angeles Seattle St. Louis are the cities with Low Price of Bread and Hamburger but high price of Apple.

Considering four cluster in K mean Cluster Analysis.

Now Cluster 4 has Anchorage and Honolulu. Cluster 1 has St Louis, Buffalo, Milwaukee And Minneapolis. Cluster 3 contains Chicago, Seattle, Houston, Washington. Cluster 2 has Boston, Atlanta, Baltimore And Kansas City

```
    Anchorage       Atlanta     Baltimore        Boston       Buffalo       Chicago    Cincinnati     Cleveland 
            4             2             2             2             1             3             2             3 
       Dallas       Detroit      Honolulu       Houston   Kansas City   Los Angeles     Milwaukee   Minneapolis 
            3             2             4             3             2             3             1             1 
     New York  Philadelphia    Pittsburgh     St. Louis     San Diego San Francisco       Seattle    Washington 
            2             2             2             1             3             3             3             3 
```

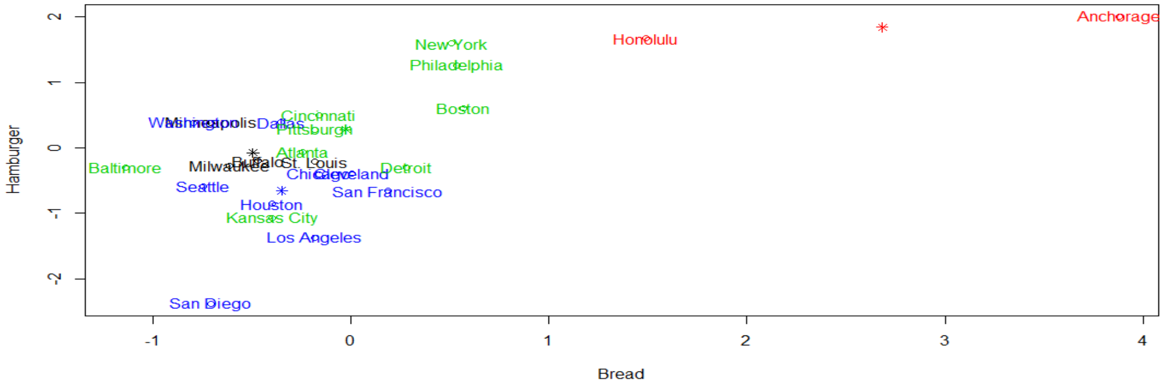
```
Cluster means:
       Bread   Hamburger        Butter     Apples     Tomato
1 -0.4950272 -0.07485785 -1.456386e+00 -0.5667544 -0.9235634
2 -0.0262288  0.27877683  3.947850e-01 -0.6903737  0.7414787
3 -0.3502707 -0.65437937 -7.526308e-05  0.6043054 -0.5485065
4  2.6843023  1.83992709  1.136578e+00  1.5208159  0.9787519
```

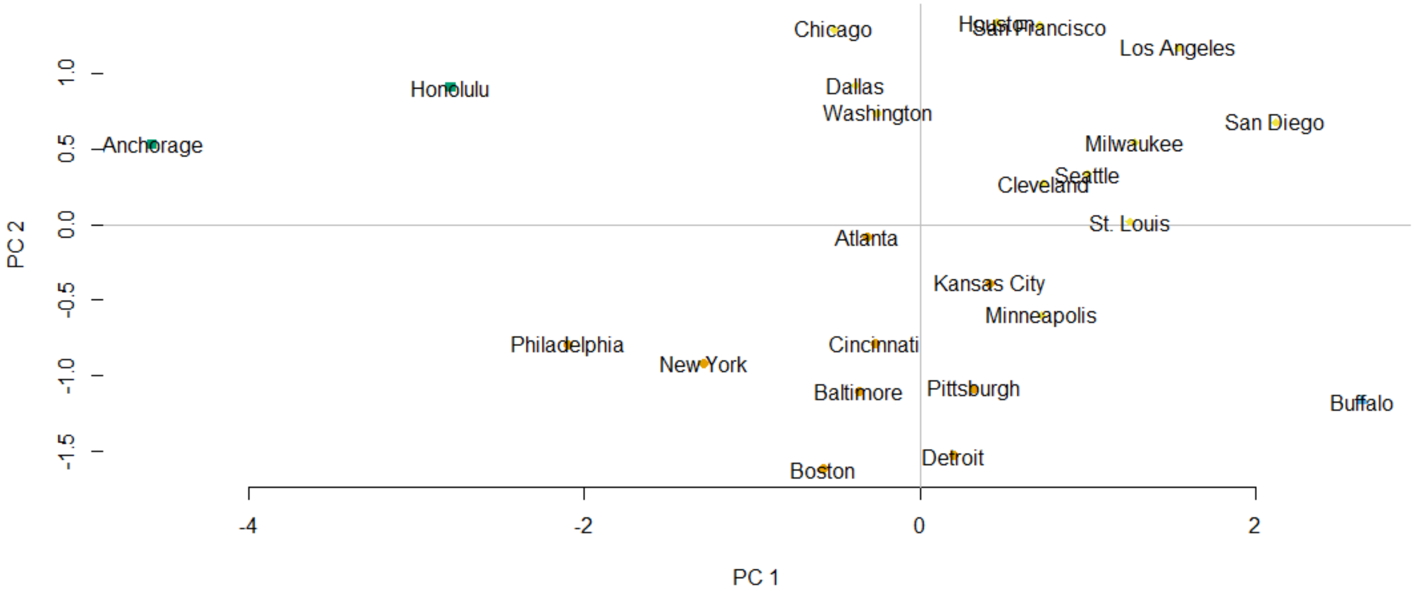
Looking at the center of each cluster. We can deduce that Cluster 1 has cities having low food price of all the food items. We can verify this with the original data

Buffalo	34.5	109.9	124.8	35.6	75.9
St. Louis	36.9	109.8	140	46.7	79

Vandana Agrawal

For both the cities the food price is comparatively low with the other cities. In cluster we have cities having low price of bread and apple. Cluster three is cities with high price of apple and low price of rest all the food items. (Chicago, Houston etc) Cluster four contains the expensive cities in terms of given food item prices. (Anchorage, Honolulu) Total Square Sum within the cluster is 47.7%. There is not much reduction in TSS from cluster 3 to 4.





Understanding the clusters using Principal component. We see that Buffalo is placed at high value of PCA1 and Very low value of PC2 it's a complete outlier and placed in One Cluster with very low food price of Apple, Hamburger and Bread. Chicago, Dallas St Louis has high PC2 value and are in cluster where Apples are expensive. New York, Boston, Detroit are cities has low PC2 and moderate PC1 meaning these cities has Low price for Apples and medium to low price of Hamburger and Bread. With the above plot it is easy to identify the clusters and distinguish the cities.