Hemachandar Nagarajan

Cluster Analysis- Food Data Set

Do the step-by-step process (listed in the R Code) and explain your answers. Do it for single link and k-means. You may pick either unstandardized or standardized data if you have a proper justification. Else, do both and state which is best after analyzing the robustness of your results

Cluster Analysis is a form of exploratory data analysis where observations are divided into meaningful groups that share common characteristics(features).

7.8 Refer to the food price data in file FOODP.DAT. Perform cluster analysis on the principal components scores to group the cities. How are the cities belonging to the same group similar, and how are they different from those belonging to other groups?

Hierarchical Clustering

Data Exploration: The data consists of prices of 5 different food items across 24 US Cities. The prices are in Cents per pound. Below table gives a summary of the data set.

```
> summary(data)
     Bread          Hamburger          Butter           Apples           Tomato      
 Min.   :28.90   Min.   : 84.5   Min.   :123.2   Min.   :35.60   Min.   : 75.90  
 1st Qu.:34.20   1st Qu.:107.0   1st Qu.:139.8   1st Qu.:46.42   1st Qu.: 84.17  
 Median :36.90   Median :109.8   Median :143.5   Median :51.10   Median : 89.90  
 Mean   :38.44   Mean   :112.2   Mean   :144.2   Mean   :51.74   Mean   : 89.76  
 3rd Qu.:40.20   3rd Qu.:117.1   3rd Qu.:150.5   3rd Qu.:58.08   3rd Qu.: 94.62  
 Max.   :70.90   Max.   :135.6   Max.   :162.3   Max.   :65.10   Max.   :104.50  
```

Should we standardize the data set or not?

Source: https://medium.com/@swethalakshmanan14/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff

https://builtin.com/data-science/when-and-why-standardize-your-data

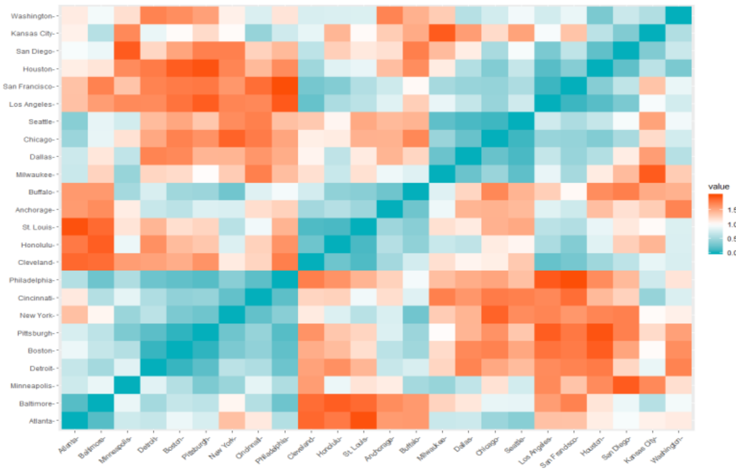
A variable that ranges between 0 and 1000 will outweigh a variable that ranges between 0 and 1 even though they are measure in the same unit. Using these variables without standardization will give the variable with the larger range weight of 1000 in the analysis. Transforming the data to comparable scales can prevent this problem. Typical data standardization procedures equalize the range and/or data variability. If we don't standardize the data the high price items will determine which cluster each city ends up in versus all of the variables.

For the above reasons we have standardized the data for the analysis.

```
> summary(data)
     Bread            Hamburger           Butter             Apples             Tomato        
 Min.   :-1.1405   Min.   :-2.3851   Min.   :-2.27736   Min.   :-1.93055   Min.   :-1.87303  
 1st Qu.:-0.5070   1st Qu.:-0.4531   1st Qu.:-0.48098   1st Qu.:-0.63554   1st Qu.:-0.75462  
 Median :-0.1843   Median :-0.2059   Median :-0.07727   Median :-0.07627   Median : 0.01915  
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000  
 3rd Qu.: 0.2102   3rd Qu.: 0.4194   3rd Qu.: 0.67868   3rd Qu.: 0.75816   3rd Qu.: 0.65776  
 Max.   : 3.8795   Max.   : 2.0076   Max.   : 1.96026   Max.   : 1.59858   Max.   : 1.99242  
```

```
> rho
              Bread Hamburger    Butter    Apples    Tomato
Bread     1.0000000 0.6490532 0.3301770 0.3187031 0.3620681
Hamburger 0.6490532 1.0000000 0.2447778 0.1908956 0.5557993
Butter    0.3301770 0.2447778 1.0000000 0.2351424 0.4361291
Apples    0.3187031 0.1908956 0.2351424 1.0000000 0.1333844
Tomato    0.3620681 0.5557993 0.4361291 0.1333844 1.0000000
```

Hemachandar Nagarajan



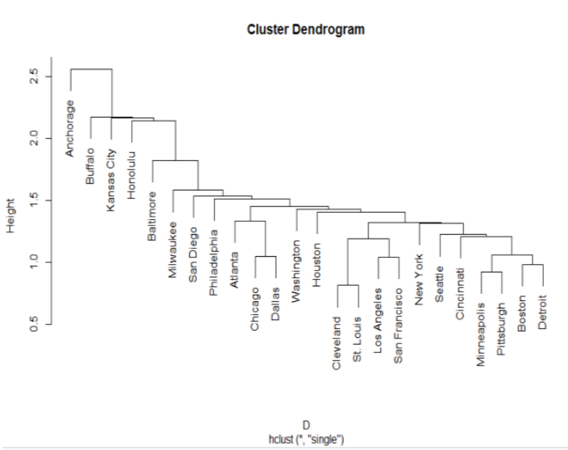
Summary and Correlation matrix after standardization.

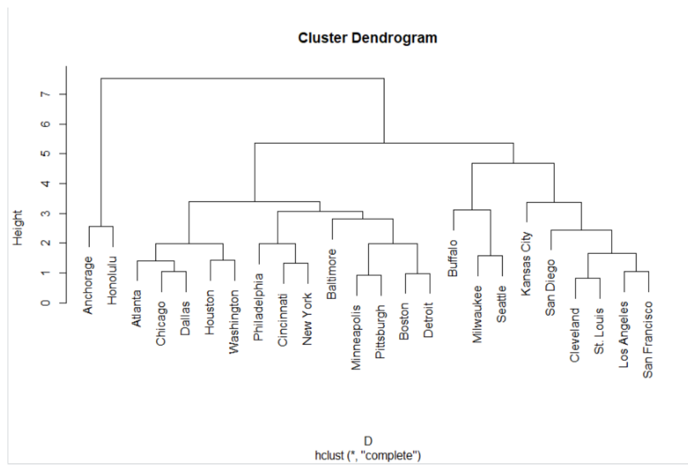
Compute a Distance Matrix

A distance matrix is calculated using R. It gives distance (Euclidean distance) between each of the observations. It helps us understand how similar or how dissimilar each observation is from other ones. [Distance = 1- similarity]. Lesser the distance more will be the similarity. As per the distance matrix, Pittsburgh and Houston have the highest distance which means that both are very dissimilar in terms of prices of the food items. Whereas, Pittsburg and Baltimore are very similar. (Euclidean distance for the standardized data)

Plot the Dendrograms

I have used both Single Linkage and Complete Linkage for the clustering. As you can see the complete linkage is more clear as it uses maximum distance between the clusters.
Complete link clusters are generally preferred over single link clusters as they tend to yield more balanced dendrograms. (ISLR, pg. 395)
Citation : https://towardsdatascience.com/introduction-hierarchical-clustering-d3066c6b560e





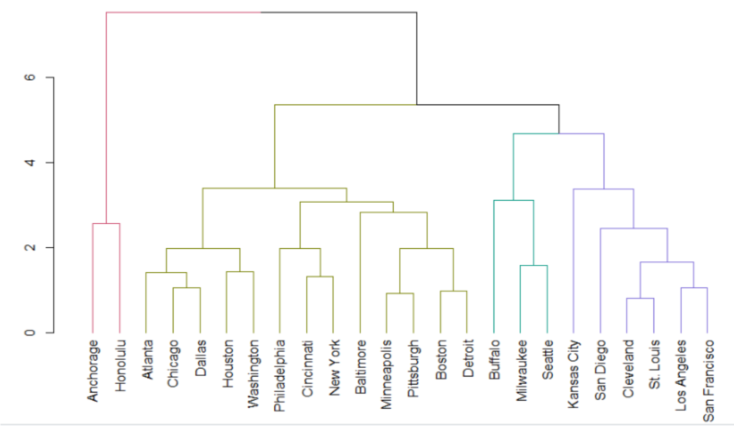
We can get an idea which cities are like each other in terms of food prices using their branches. The height of the branch gives their distance.

Choosing a cluster: We can cut the dendrogram like a tree to identify the number of clusters. As you can see from the Complete Linkage Dendrogram we can group the cities into 4 clusters as shown below.

	cluster	n
	<int>	<int>
1	1	2
2	2	13
3	3	3
4	4	6

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

Hemachandar Nagarajan



	Bread	Hamburger	Butter	Apples	Tomato	cluster
1	3.879538966	2.0075514	1.16909152	1.45501855	1.39773432	1
2	-0.244027483	-0.0641127	0.00943799	0.25870320	0.83008074	2
3	-1.140454972	-0.2962078	0.73557618	-0.50693863	1.99241903	2
4	0.568733440	0.6063844	-0.23983333	-1.27258046	0.91117411	2
5	-0.471122447	-0.2016506	-2.10394929	-1.93055390	-1.87303157	3
6	-0.160360917	-0.4079574	0.12865471	1.59857639	0.60031619	2
7	-0.160360917	0.5032310	0.58384581	-0.73423855	0.14078710	2
8	0.006972214	-0.3907651	-0.16396815	-0.17197033	-0.88639558	4
9	-0.351598782	0.3914815	-0.18564391	1.27557125	0.12727154	2
10	0.281876644	-0.2962078	-0.44575312	-1.44006461	0.85711186	2
11	1.489065662	1.6723028	1.10406422	1.58661324	0.55976951	1
12	-0.399408248	-0.8549554	0.65971100	0.90471349	-0.71069328	2
13	-0.399408248	-1.0698583	1.96025702	-1.09313315	-0.25116418	4
14	-0.184265650	-1.3793185	-0.41323947	0.35440842	-1.41350247	4
15	-0.614550845	-0.2704195	-2.27735542	0.71330303	-0.27819531	3
16	-0.710169777	0.3828854	-0.98764729	-0.44712286	-0.08897745	2
17	0.508971608	1.5949378	0.48630486	-0.49497548	0.31648940	2
18	0.532876341	1.2596892	1.03903692	0.01944012	1.58695218	2
19	-0.184265650	0.2711358	-0.57580772	-0.94957531	0.28945828	2
20	-0.184265650	-0.2102467	-0.45659100	-0.60264386	-1.45404916	4
21	-0.710169777	-2.3850641	0.18284413	-0.38730710	-1.00803563	4
22	0.186257712	-0.6572447	-0.55413195	0.89275033	-1.06209788	4
23	-0.746026877	-0.5884758	-0.80340327	0.27066635	-0.15655525	3
24	-0.793836343	0.3828854	1.14849954	0.70133988	-0.42686648	2

Cluster centers:

	cluster	Bread	Hamburger	Butter	Apples	Tomato
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	2.68	1.84	1.14	1.52	0.979
2	2	-0.173	0.267	0.181	-0.0836	0.494
3	3	-0.611	-0.354	-1.73	-0.316	-0.769
4	4	-0.214	-1.02	0.0925	-0.168	-1.01

From the cluster centers table you can see that Cluster 1 (Anchorage and Honolulu) has the highest food price range of all the cities. It aligns with actual situation as well, as these cities are located far from the mainland the prices tend be high. Cluster 2, which include cities highlighted in green above have average hamburger, butter and tomato prices and below average bread and apples. Cluster3 (Buffalo, Milwaukee and Seattle) have the lowest butter price among all the cities. The other food prices are also low in this cluster. Cluster 4 has very low Hamburger price and a average butter price. The rest of the food items are cheap in this cluster also.

Total sum of squares from cluster centroid

Cluster	Sum of Squares
1	5.408
2	39.847
3	8.8
4	21.44
TSS	75.495

TSS is the sum of the squares of the differences between the dependent variable and its mean.

Hemachandar Nagarajan

K-means Clustering

k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, The key idea of K-Means clustering is that it tries to minimize the within cluster variation. The best model for a given k, is the one which minimizes the total within cluster sum of squares. The best model minimizes the amount of variance within clusters and maximizes the variance between clusters.

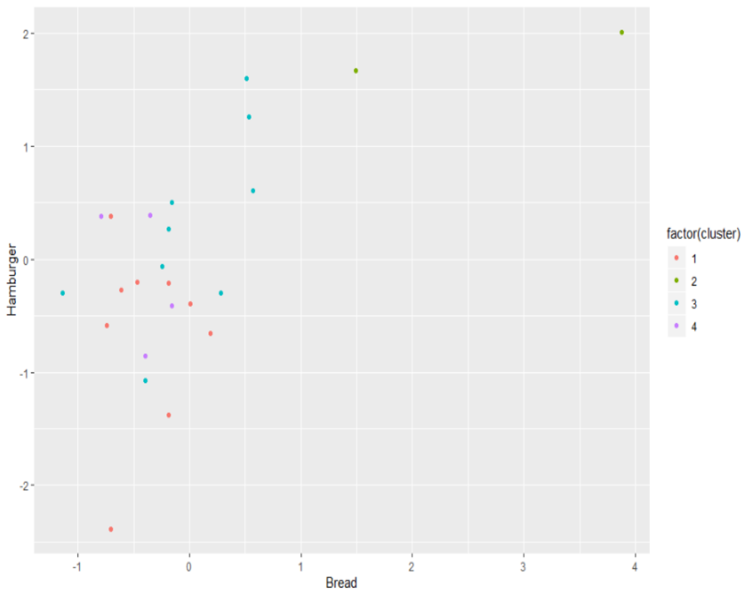
```
K-means clustering with 4 clusters of sizes 9, 2, 9, 4

Cluster means:
       Bread  Hamburger     Butter     Apples     Tomato
1 -0.3808157 -0.6333666 -0.8419380 -0.1453855 -0.9134267
2  2.6843023  1.8399271  1.1365779  1.5208159  0.9787519
3 -0.0262288  0.2787768  0.3947850 -0.6903737  0.7414787
4 -0.4263011 -0.1221365  0.4378053  1.1200502 -0.1024930

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

Within cluster sum of squares by cluster:
[1] 20.714527  3.275242 19.864302  3.884990
 (between_SS / total_SS =  58.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"
> cl$tot.withinss
[1] 47.73906
```

The number of clusters (k) were taken as 4 as we took in the Hierarchical cluster. The best model with less total within cluster sum of squares is got using multiple simulation.
As k-means does not give us a unique output we can run the code multiple times to find the best model. Alternatively, we can set the number of iterations in the kmeans function in R. The best model for a given k, is the one which minimizes the total within cluster sum of squares. In other words, it minimizes the amount of variance within clusters and maximizes the variance between clusters. The total within sum of squares is got 47.8 which is much less than we got from hierarchical clustering (75.495).

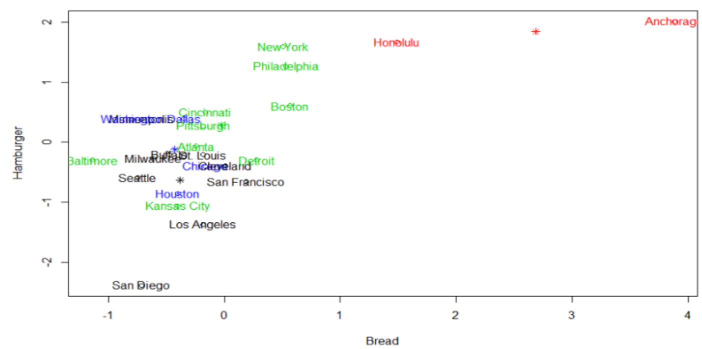
```
> segment_k
          Bread  Hamburger      Butter      Apples      Tomato cluster
1   3.879538966  2.0075514  1.16909152  1.45501855  1.39773432       2
2  -0.244027483 -0.0641127  0.00943799  0.25870320  0.83008074       3
3  -1.140454972 -0.2962078  0.73557618 -0.50693863  1.99241903       3
4   0.568733440  0.6063844 -0.23983333 -1.27258046  0.91117411       3
5  -0.471122447 -0.2016506 -2.10394929 -1.93055390 -1.87303157       1
6  -0.160360917 -0.4079574  0.12865471  1.59857639  0.60031619       4
7  -0.160360917  0.5032310  0.58384581 -0.73423855  0.14078710       3
8   0.006972214 -0.3907651 -0.16396815 -0.17197033 -0.88639558       1
9  -0.351598782  0.3914815 -0.18564391  1.27557125  0.12727154       4
10  0.281876644 -0.2962078 -0.44575312 -1.44006461  0.85711186       3
11  1.489065662  1.6723028  1.10406422  1.58661324  0.55976951       2
12 -0.399408248 -0.8549554  0.65971100  0.90471349 -0.71069328       4
13 -0.399408248 -1.0698583  1.96025702 -1.09313315 -0.25116418       3
14 -0.184265650 -1.3793185 -0.41323947  0.35440842 -1.41350247       1
15 -0.614550845 -0.2704195 -2.27735542  0.71330303 -0.27819531       1
16 -0.710169777  0.3828854 -0.98764729 -0.44712286 -0.08897745       1
17  0.508971608  1.5949378  0.48630486 -0.49497548  0.31648940       3
18  0.532876341  1.2596892  1.03903692  0.01944012  1.58695218       3
19 -0.184265650  0.2711358 -0.57580772 -0.94957531  0.28945828       3
20 -0.184265650 -0.2102467 -0.45659100 -0.60264386 -1.45404916       1
21 -0.710169777 -2.3850641  0.18284413 -0.38730710 -1.00803563       1
22  0.186257712 -0.6572447 -0.55413195  0.89275033 -1.06209788       1
23 -0.746026877 -0.5884758 -0.80340327  0.27066635 -0.15655525       1
24 -0.793836343  0.3828854  1.14849954  0.70133988 -0.42686648       4
```

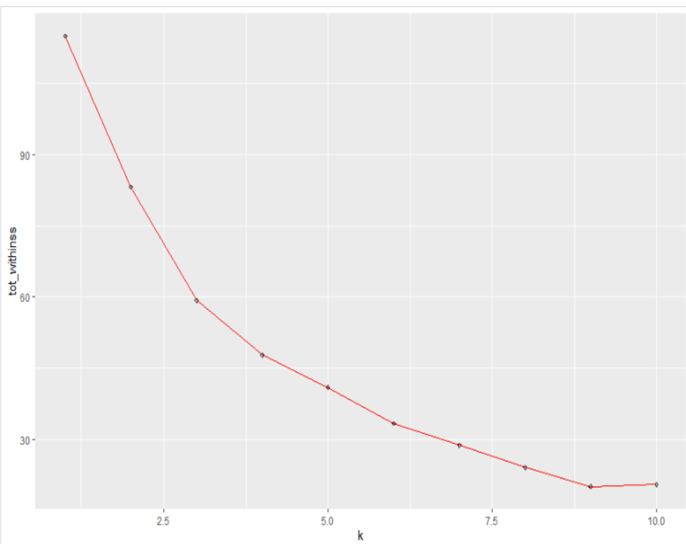
Plotting the cluster: Using K= 4 we cluster the cities as shown. As the data is 5D we cannot plot and show the clusters clearly. I have plotted a 2D plot between bread and hamburger as shown below.

Hemachandar Nagarajan



<u>Interpretation:</u> The cities falling in Cluster 2 (green) have high bread and hamburger prices while the cities in Cluster 1 (red) have fairly low prices of bread and hamburger. The other plot with city names will give a clearer idea on bread and hamburger in each cities.





<u>Selection of K</u>: You can then determine the optimal *k* clusters by using something called the **elbow** method (Scree Plot). You want to find the point of diminishing returns when selecting a range of clusters. You can do this by plotting the number of clusters on the X-axis and the inertia (within-cluster sum-of-squares criterion) on the Y-axis. You then select *k* for which you find a bend. For the plot you can see that a bend is there k=4.

<u>Interpretation of Clusters:</u>

	Bread	Hamburger	Butter	Apples	Tomato
1	-0.3808157	-0.6333666	-0.8419380	-0.1453855	-0.9134267
2	2.6843023	1.8399271	1.1365779	1.5208159	0.9787519
3	-0.0262288	0.2787768	0.3947850	-0.6903737	0.7414787
4	-0.4263011	-0.1221365	0.4378053	1.1200502	-0.1024930

Cluster 2 which consist of Anchorage and Honolulu have high food prices. On the contrary cities in Cluster 1 have low food prices. In Cluster 4, which consists of cities Chicago, Houston and Washington have very high prices of apples. Cluster 3 cities have average prices for Hamburger, Butter and Tomato while Bread and Apples are cheaper there.

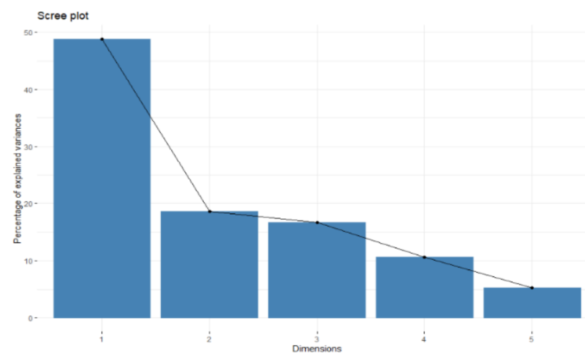
Hemachandar Nagarajan

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

Comparison of K means and Hierarchical Clustering: From the above two clustering Anchorage and Honolulu moved from Cluster 1 to Cluster 2. Buffalo, Milwaukee and Seattle have moved from Cluster 3 to Cluster 1. Similarly, some more cities have shifted their clusters and the changes in the food prices can be seen in the above Cluster centers Matrix.

<u>Plot Clusters With PCA</u>:

```
Standard deviations (1, .., p=5):
[1] 1.5618436 0.9641531 0.9128186 0.7299813 0.5147260

Rotation (n x k) = (5 x 5):
                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
```



The first 4 Principal Components explain 92% of variance in the data set.

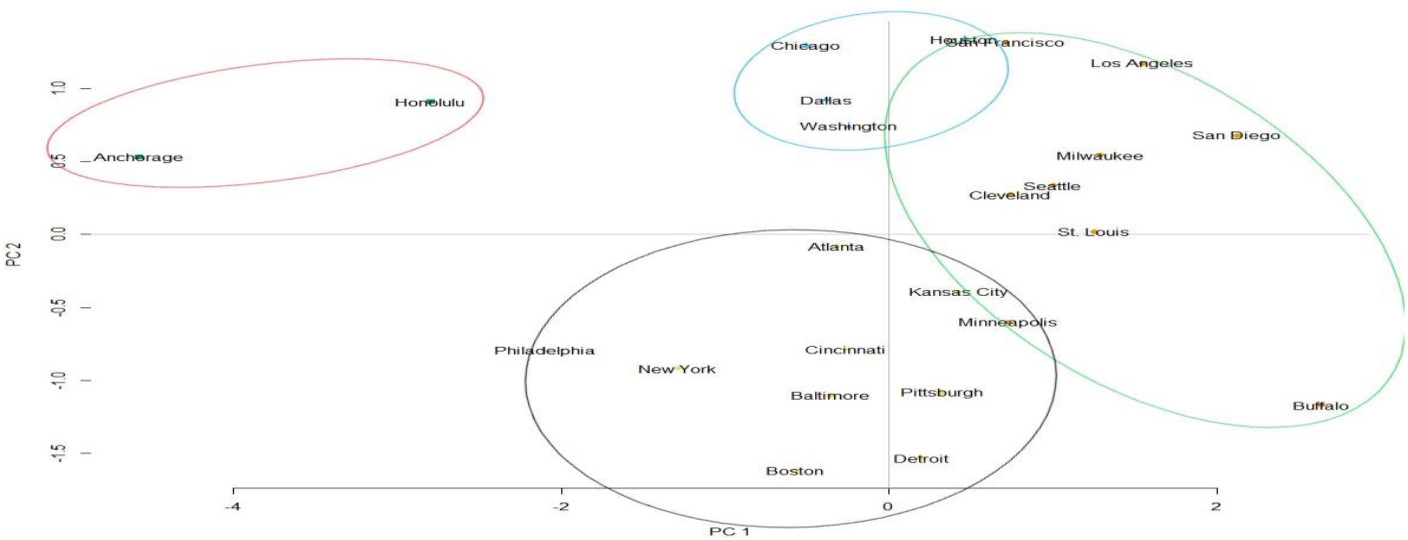
The angle between the arrows representing the food items can be interpreted as the correlation between them. (small angle = High +ve Correlation)

The angle between the food item and the PCA Axis can be interpreted as the correlation between the two (Small angle = High +ve Correlation)

The length of the arrows is proportional to the Standard Deviation of the food items.

(Source: Data Camp)

Hemachandar Nagarajan



.

Interpretation: Anchorage and Honolulu have a low value of PC1 and high value of PC2 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 (PC2) and average price of Hamburger and bread. Green Cluster with (Los Angeles Seattle, St. Louis, Cleveland) are the cities with low price of Bread and Hamburger but high price of Apple.