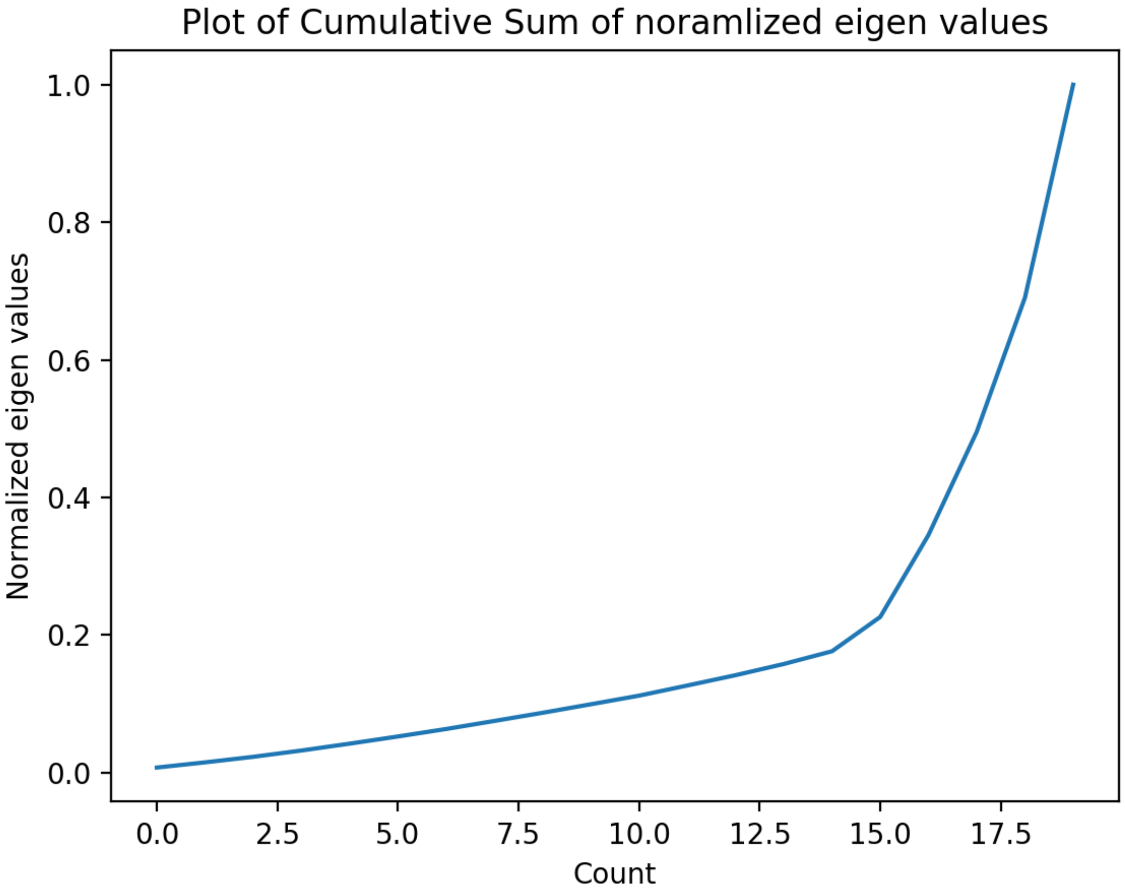
Pratishta Rao, Srikant Lakshminarayan

Big Data Analytics - Home Work 9

0. Plot the cumulative sum of these normalized eigenvalues.

Solution:

Plot for the cumulative sum of these normalized Eigen values is follows:



Pratishta Rao, Srikant Lakshminarayan

1. Record and report the size of the last 20 smallest clusters merged.

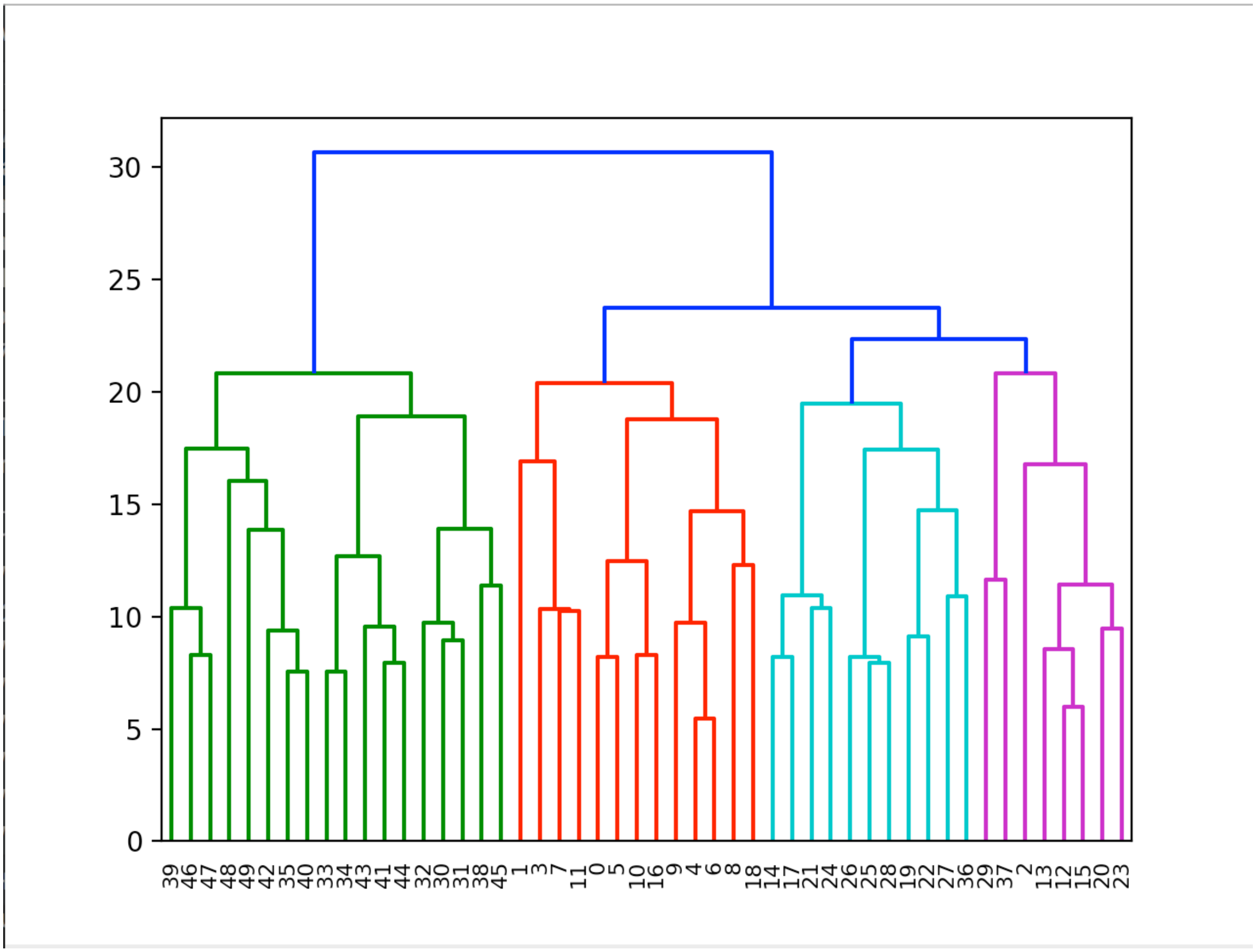
Solution:

The size of the last 20 smallest clusters merged are as follows:

114, 155, 115, 116, 144, 156, 135, 157, 158, 145, 146, 117, 147, 159, 160, 161, 148, 149, 133, 162

2. Can you support this guess with a dendrogram?

Solution:



3. Print out the first four eigenvectors to one significant digit. What does this tell you about the attributes? Which attribute is least important? Which attribute is most important?

Solution:

The first four eigen vectors associated with the four highest eigen values:

```
[[ 0.3 -0.1  0.4  0.1  0.2  0.   0.1 -0.2  0.3 -0.2 -0.2  0.1 -0.1  0.3
   0.1 -0.1  0.5 -0.1  0.1  0.3]
 [0.   0.1  0.1 -0.3 -0.2 -0.1 -0.3 -0.1 0.   0.  -0.6 -0.2  0.2 -0.2
   0.1  0.1  0.2  0.4  0.   0.1]
 [-0.1  0.2  0.3 -0.4  0.   0.2  0.2  0.   0.   0.   0.3  0.4  0.2 -0.1
   0.  -0.2 -0.1  0.3  0.2  0.3]
 [0.  0.   0.   0.   0.1  0.8 -0.6 -0.1  0.   0.1  0.2 -0.1 0.   0.
   0.2  0.1  0.1 -0.   0.1 -0. ]]
```

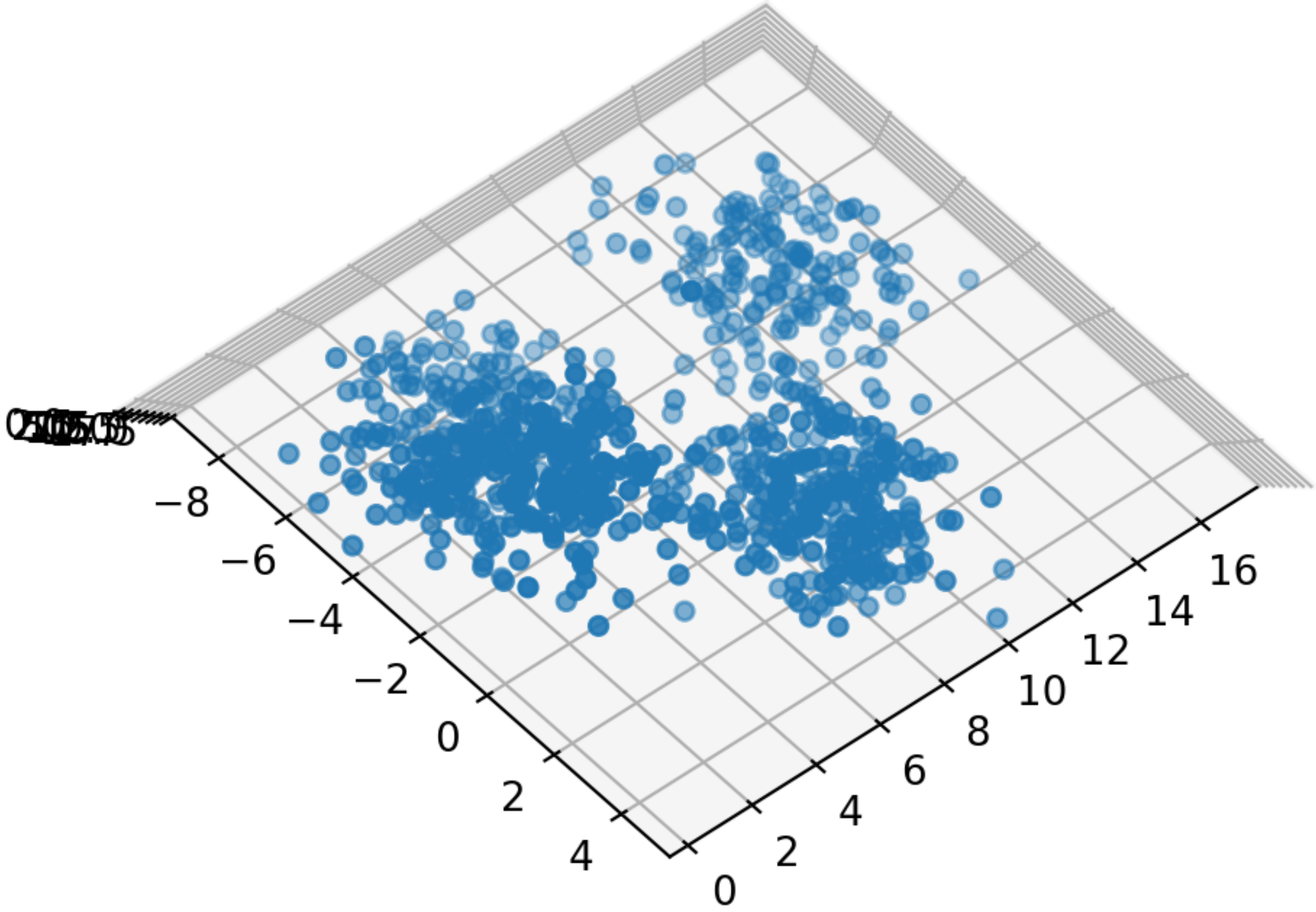
If we separate each of the first 4 values from Eigen vectors and assign each attributes these 4 values, each attributes has 4 values and by looking at number of zero and non zero values we get worst and best attributes.

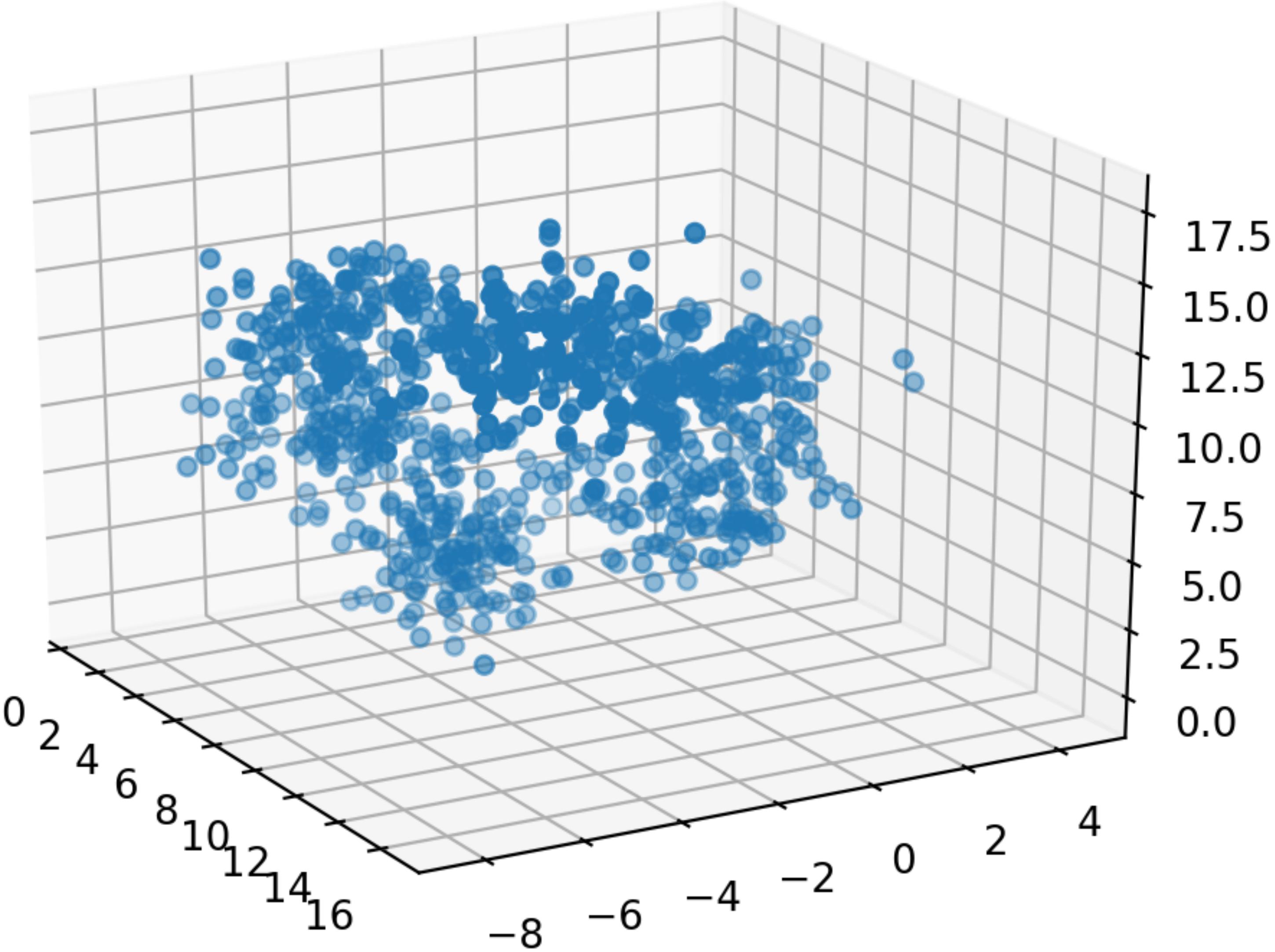
```
Beans 0.3 -0.1 0.4 0.1
Bread 0 0.1 0.1 -0.3
Cereal -0.1 0.2 0.3 -0.4
ChdBby 0 0 0 0
Chips 0.1 0.6 0.1 0
Corn 0.4 0.1 0 0
Eggs 0 0.1 -0.4 -0.1
Fish -0.1 -0.1 -0.2 -0.1
Fruit 0.2 -0.1 -0.2 0.1
Meat 0 0.2 -0.4 0
Milk 0.1 0 0 -0.6
Pepper 0.4 0.2 0 -0.2
Rice 0.2 -0.3 -0.1 -0.2
Salsa 0.3 0.2 0.3 -0.1
Sauce -0.1 -0.3 0.4 0.1
Soda -0.2 0.3 0 0.3
Tomato 0.3 0.1 -0.2 0.2
Tortya 0.4 0.2 0 0.2
```

Vegges 0.2 -0.3 -0.2 -0.1
Yogchs 0.3 -0.2 0 0

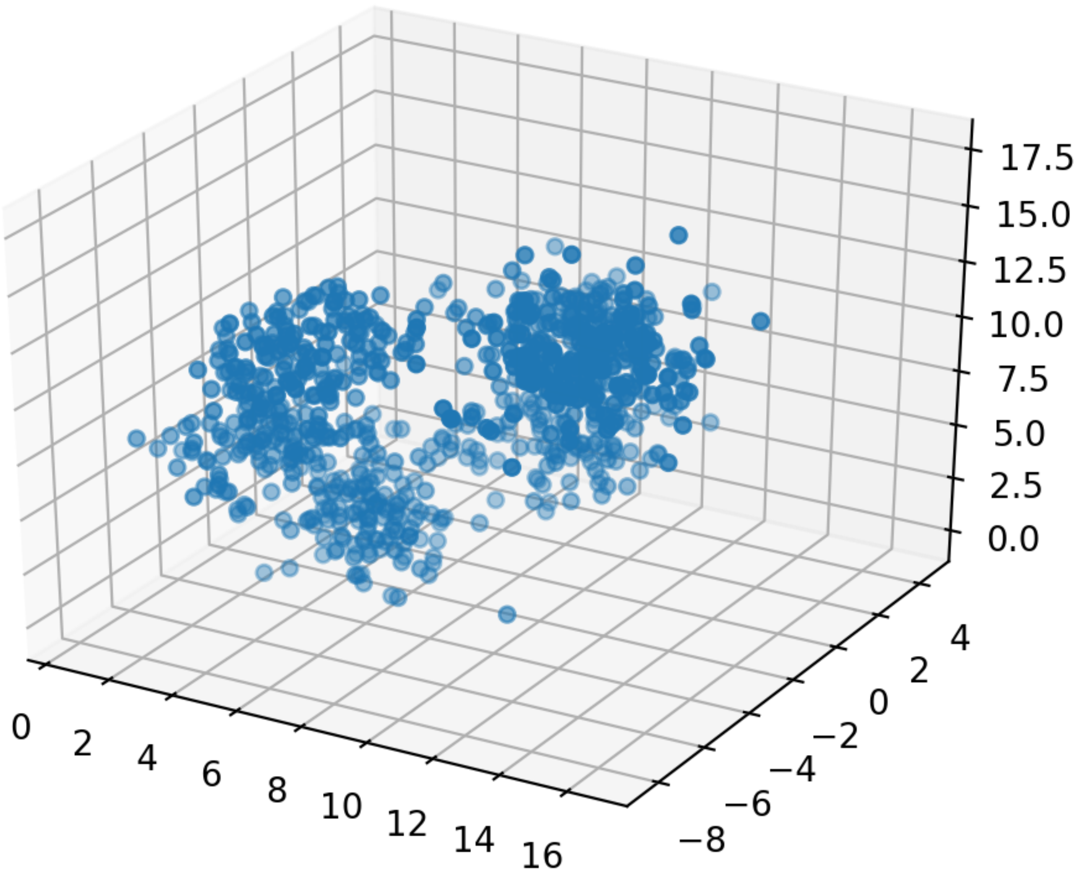
The least important attribute is : ChdBby as the value of the vector is zero.
The most important attribute is : Chips and Tortya as the value of these attributes is highest.

4. . PLOT: Generate a 3D plot of these projected points, (use the first three of the four axes) and show a scatter gram of this data in 3D.

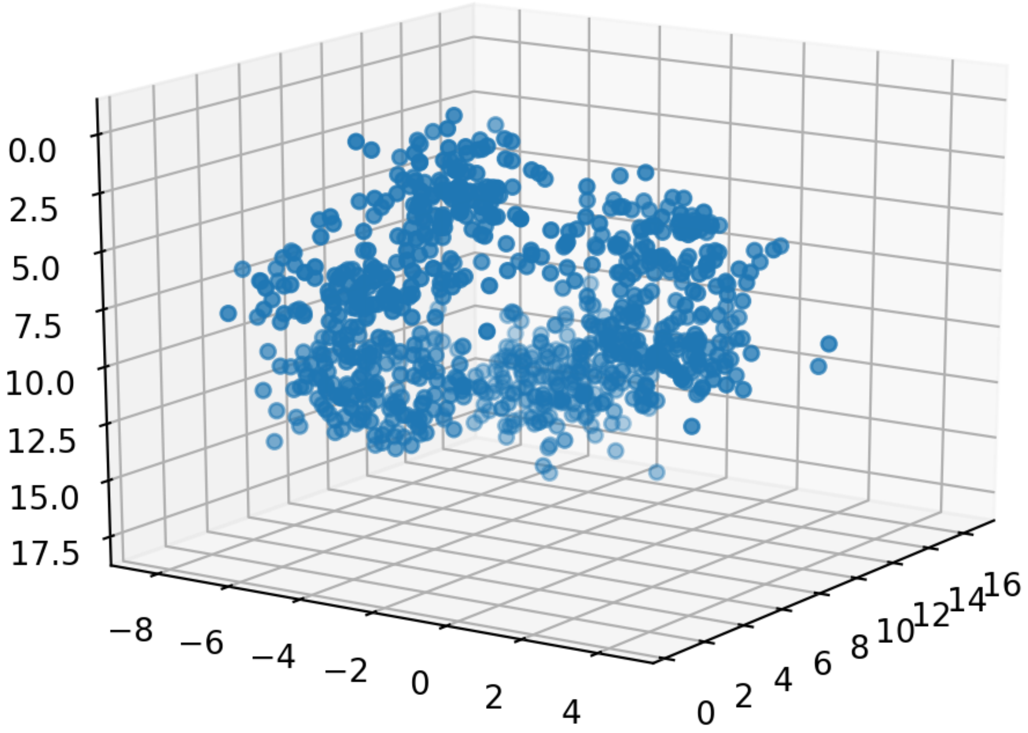




Solution: 3D plot obtained is as follows:



0
2
4
6
8
10
12
14
16
−8
−6
−4
−2
0
2
4
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

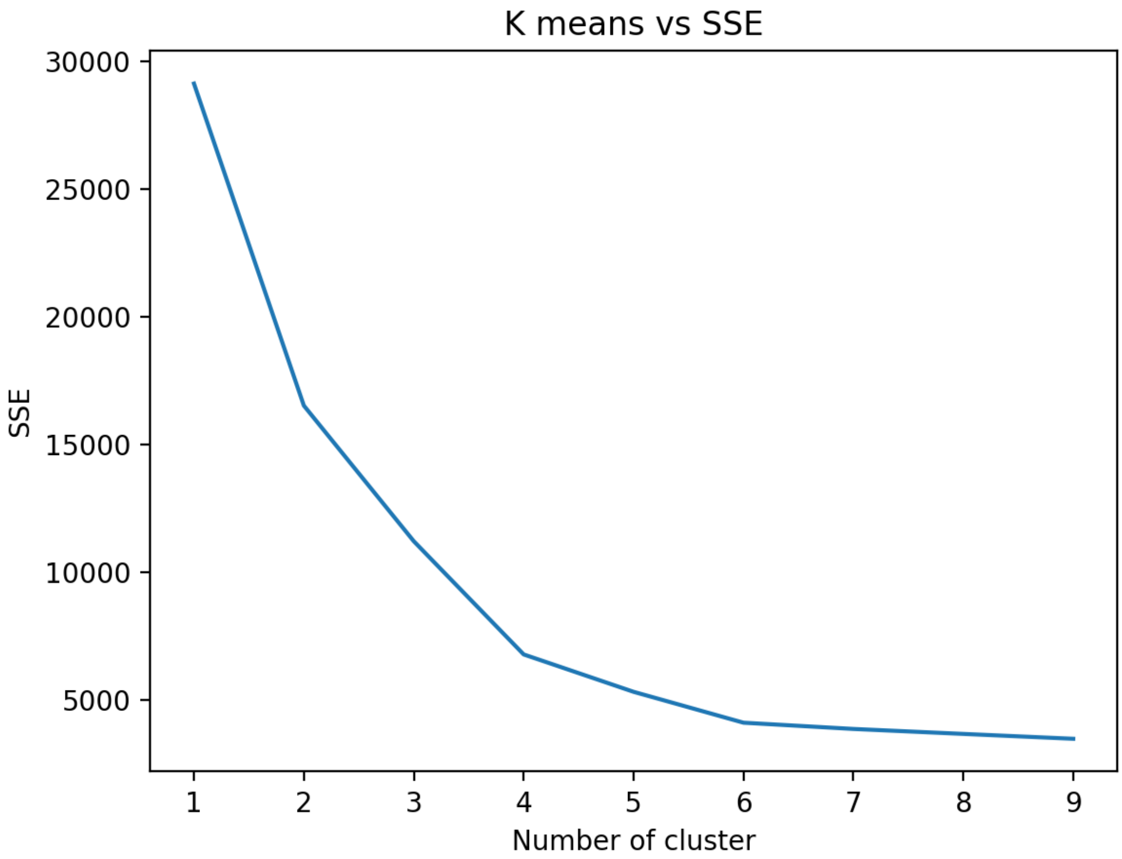


0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
−8
−6
−4
−2
0
2
4
0
2
4
6
8
10
12
14
16

5. Plot the SSE (or avg SSE per cluster) versus K.

Solution:

Plot for K vs SSE is as follows:



6. Using this plot, how many clusters would you suspect from K-Means?

Solution:

From the plot above, we think that k means has 6 clusters.

7. By inspection, of your 3D plot, how many clusters do you suspect?

Solution:

From the 3D plot, we suspect it to have 6 clusters

8. When you have clustered to six clusters, report the size of each cluster, from low to high.

Solution:

For K-means cluster size for k = 6 is as follows:

```
Cluster Size:
[143 128 160 117 148 120]

Sorted cluster size:
[117 120 128 143 148 160]
```

For agglomeration, cluster size is as follows:

```
[115 118 134 136 150 163]
```

9. When you have clustered to six clusters, report the average prototype of these six clusters.

Solution:

Average prototype for k = 6

```
[[ 3.64416667e+00 -4.64166667e+00  6.77833333e+00]
 [ 1.32425000e+01 -4.77187500e+00  1.36231250e+01]
 [ 5.63828125e+00 -3.13359375e+00  2.77734375e+00]
 [ 9.15042735e+00  4.44444444e-01  5.35299145e+00]
 [ 9.78175676e+00 -4.05405405e-03  1.02189189e+01]
 [ 5.44195804e+00 -5.25244755e+00  1.17811189e+01]]
```

10. What typifies each of the six clusters? What name should we give each prototype?

Solution:

The name given for each prototype

```
shoppers = ['Vegan', 'Kosher', 'Weed eater', 'Hillbilly', 'Family', 'Other']
```

11. You are provided with 20 shoppers to classify. Create a 1-NN classifier. **Using 1-NN**, classify each of these 20 shoppers as one of the six cluster prototypes you assigned. Which of the six nicknames goes with each of these 20 shoppers?

Solution:

The six nicknames goes with each of these 20 shoppers are as follows:

Shopper 1 Goes Into Cluster Weed eater
Shopper 2 Goes Into Cluster Other
Shopper 3 Goes Into Cluster Kosher
Shopper 4 Goes Into Cluster Hillbilly
Shopper 5 Goes Into Cluster Vegan
Shopper 6 Goes Into Cluster Hillbilly
Shopper 7 Goes Into Cluster Weed eater
Shopper 8 Goes Into Cluster Weed eater
Shopper 9 Goes Into Cluster Hillbilly
Shopper 10 Goes Into Cluster Other
Shopper 11 Goes Into Cluster Vegan
Shopper 12 Goes Into Cluster Weed eater
Shopper 13 Goes Into Cluster Other
Shopper 14 Goes Into Cluster Vegan
Shopper 15 Goes Into Cluster Kosher
Shopper 16 Goes Into Cluster Hillbilly
Shopper 17 Goes Into Cluster Vegan
Shopper 18 Goes Into Cluster Family
Shopper 19 Goes Into Cluster Weed eater
Shopper 20 Goes Into Cluster Kosher

12. What advantage is there in performing k-means clustering on data which has been projected using PCA?

Solution:

Since PCA reduces the number of dimensions of the given data set, implementing k-means is much easier and has a faster runtime as the data has lower dimensions.

13. Write a general conclusion about what you learned overall.

Solution:

The following are what we learned from this homework:

PCA reduces number of attributes i.e. removes attributes which have low variance. Trying to figure out the most important attribute and the least important attribute based on the Eigen vectors. We can also see that k-means runs faster on data with lower dimensions. By both the agglomeration methods we get the same number of clusters but k-means is much faster. Plotting a 3D data was something both of us hadn't worked on before. Overall, this was a very interesting assignment to work on.