Principles of Data Mining (420)
Spring 2022, Homework 07

Mohammed Mehboob
(mm2260@rit.edu)

March 23, 2022

Question-1: Why do we use 10 visits instead of just keeping records of every single visit?

⇒ We use 10 visits instead of keeping records for every visit for every shopper as a noise-reduction measure.

Question-4: Plot the cumulative sum of the normalized eigenvalues.

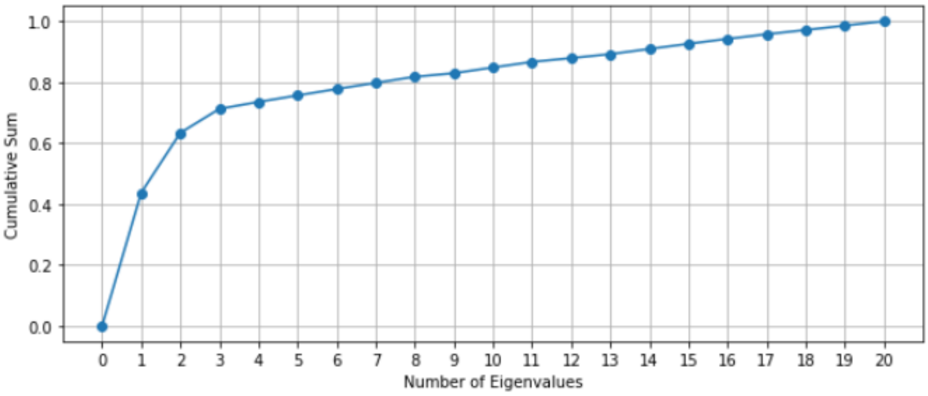
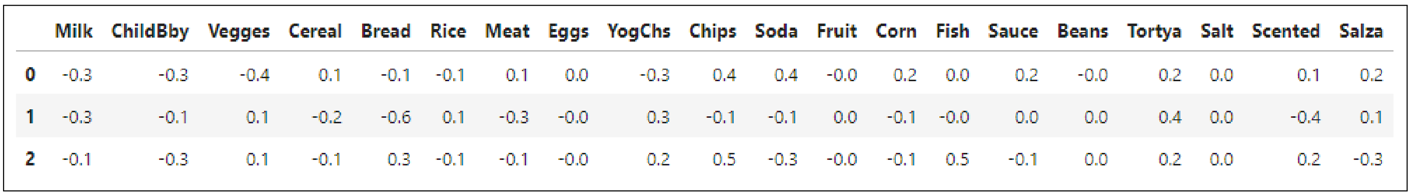


Figure 1: Cumulative Sum of Normalized Eigenvalues

Question-5: Print out the first three eigenvectors. Why does this tell you about the attributes? Which attributes are most important? Which can be ignored? Justify your answers.



	Milk	ChildBby	Vegges	Cereal	Bread	Rice	Meat	Eggs	YogChs	Chips	Soda	Fruit	Corn	Fish	Sauce	Beans	Tortya	Salt	Scented	Salza
0	-0.3	-0.3	-0.4	0.1	-0.1	-0.1	0.1	0.0	-0.3	0.4	0.4	-0.0	0.2	0.0	0.2	-0.0	0.2	0.0	0.1	0.2
1	-0.3	-0.1	0.1	-0.2	-0.6	0.1	-0.3	-0.0	0.3	-0.1	-0.1	0.0	-0.1	-0.0	0.0	0.0	0.4	0.0	-0.4	0.1
2	-0.1	-0.3	0.1	-0.1	0.3	-0.1	-0.1	-0.0	0.2	0.5	-0.3	-0.0	-0.1	0.5	-0.1	0.0	0.2	0.0	0.2	-0.3

Figure 2: First three Eigenvectors (rounded), sorted by their Eigenvalues

⇒ When we rank the eigenvectors by their corresponding eigenvalues in descending order, the Principal Component corresponding to those eigenvectors will hold more significance, since they account for more of the variance of the data.

And since the principal components are abstract entities, as linear combinations of the original independent variables, the eigenvectors for the top principal components tell us about which attributes are important to us by analyzing which attributes weigh most within them.

$$\%\ Variance = \frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

Thus the eigenvectors holding more % variance will be more useful to us for the purposes of dimensionality reduction.

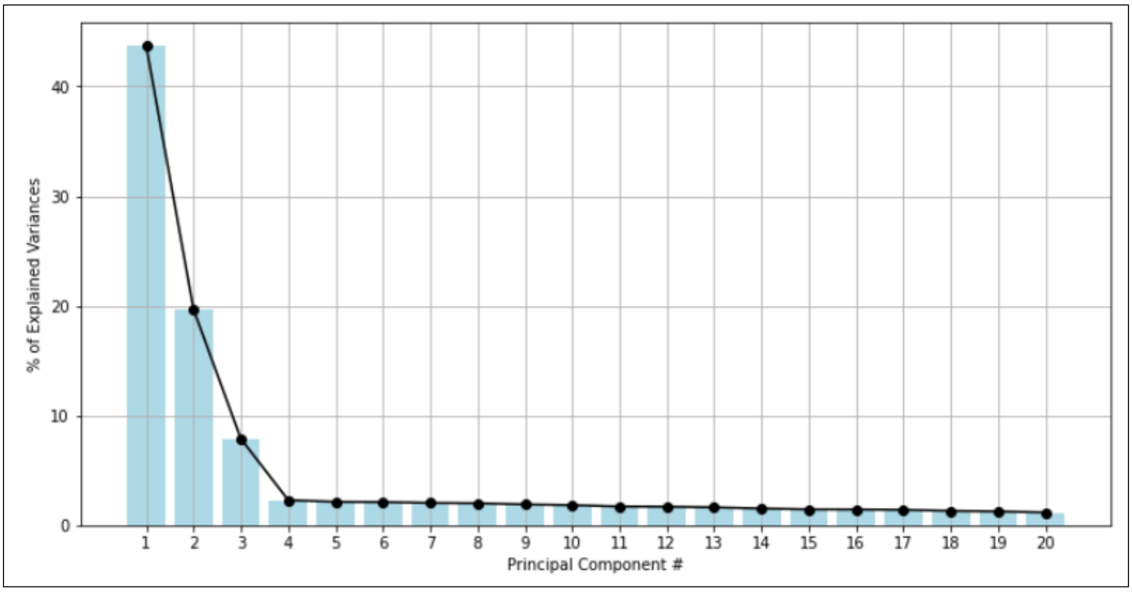


Figure 3: Percent of Explained Variances

As we can see in Figure-3, the percent of explained variance suddenly drops after the third Principal Component. Which tells us that the first three principal components are all we will need to effectively understand the data.

Therefore, attributes like Veggies, Chips or Fish seems to be important, whereas attributes like Eggs, Fruit, Beans and Salt can be completely ignored since they contribute nothing to the principal components.

Question-6: Generate a 2D plot of these projected points, and show a scatter gram of this 2D plot of points.

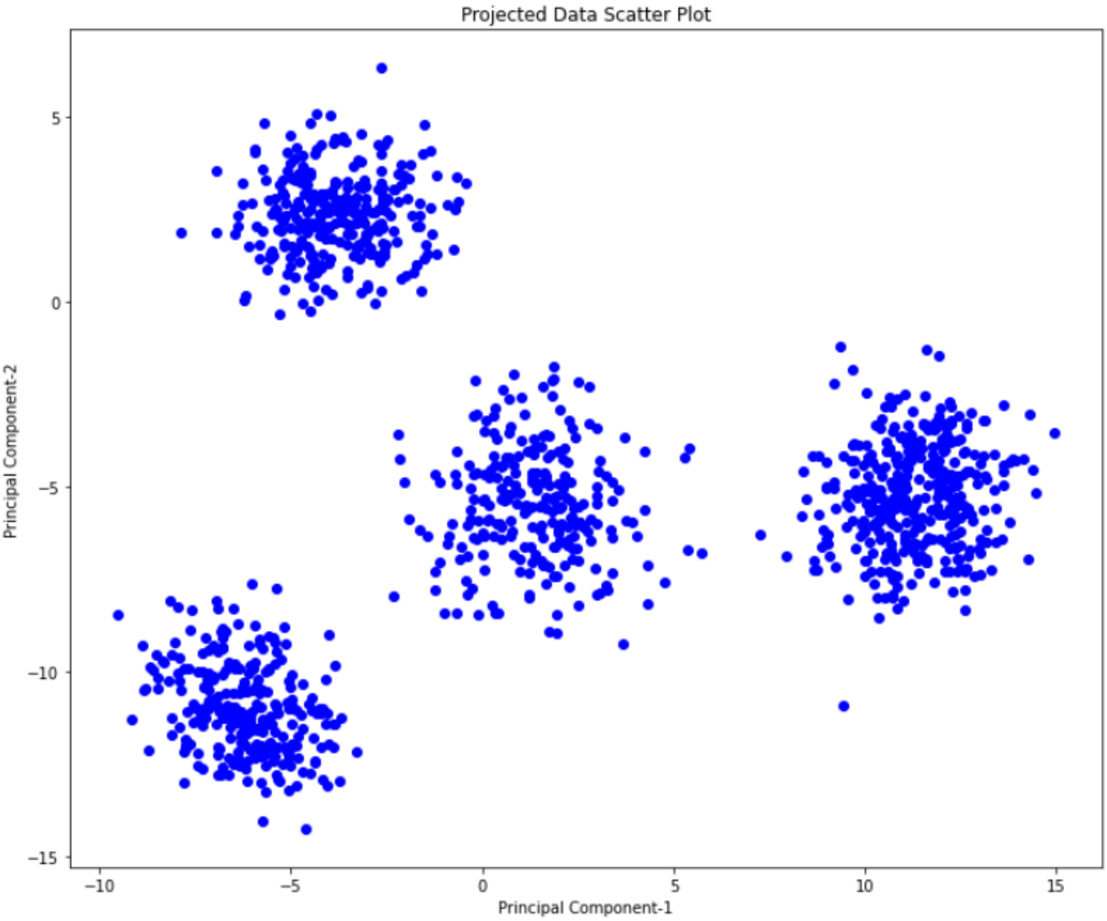


Figure 4: Data Projected Onto Principal Componenets 1 and 2

In 2-D, the scatter plot appears to consist of **4 clusters**, however, if we project the agglomeration data onto the first three principal components, the *3-D scatter plot appears to have 3 clusters*, which matches the number of clusters we saw in the last assignment (agglomeration clustering).

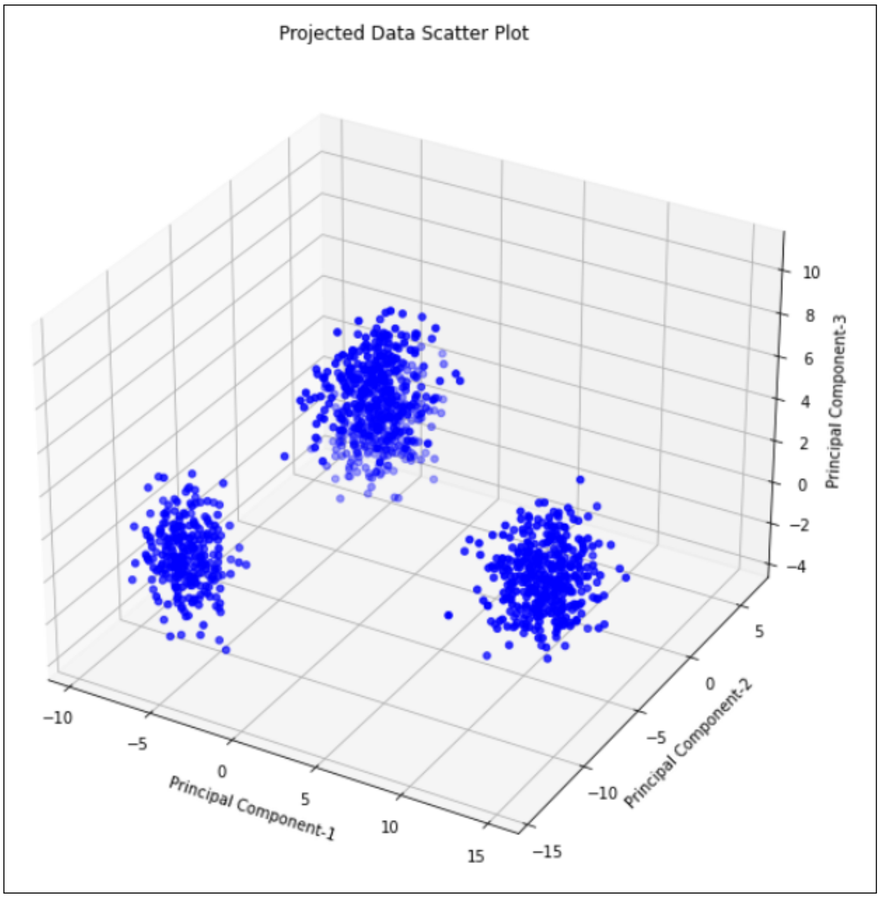
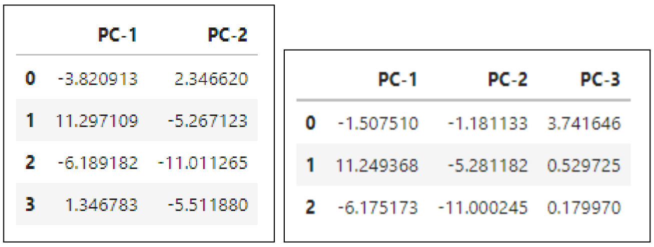


Figure 5: Data Projected Onto Principal Componenets 1, 2 and 3

Question-8: Print the center of mass vectors in PCA space.

⇒ Using the k-means algorithm to find the centers for each of the clusters, we get the following vectors as the cluster centers:

(k = 4 for 2D data, k = 3 for 3D data)



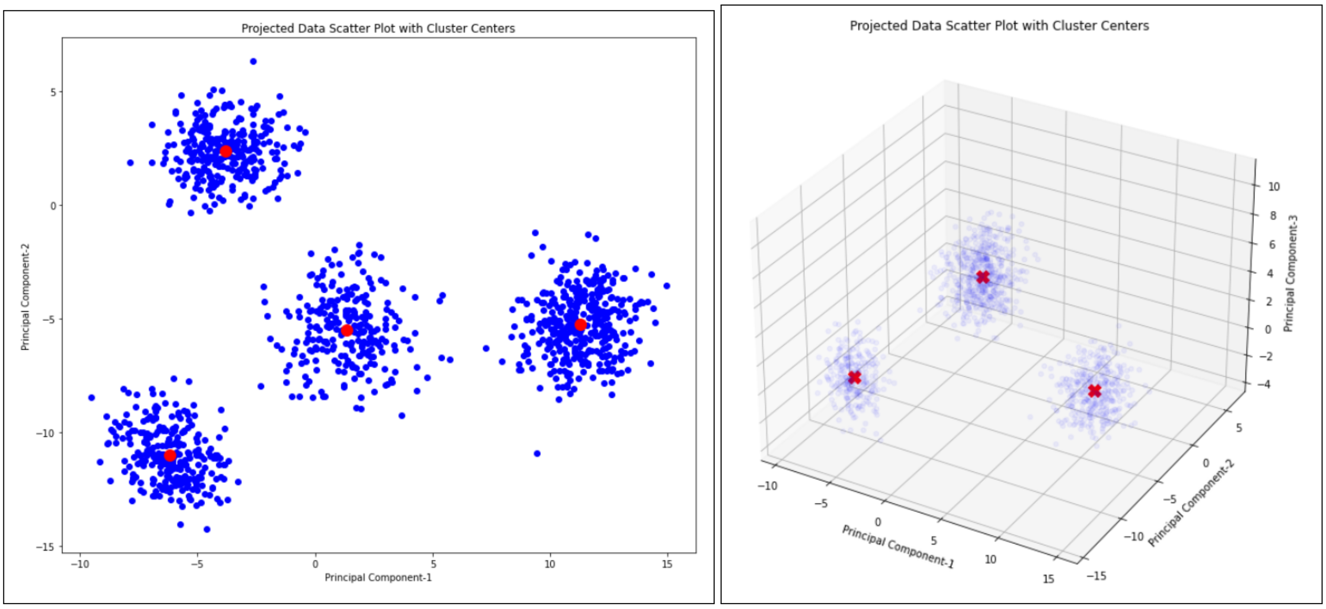
	PC-1	PC-2
0	-3.820913	2.346620
1	11.297109	-5.267123
2	-6.189182	-11.011265
3	1.346783	-5.511880

(a) Cluster Centers (2D)

	PC-1	PC-2	PC-3
0	-1.507510	-1.181133	3.741646
1	11.249368	-5.281182	0.529725
2	-6.175173	-11.000245	0.179970

(b) Cluster Centers (3D)

Figure 6: Cluster Centers (Vectors)

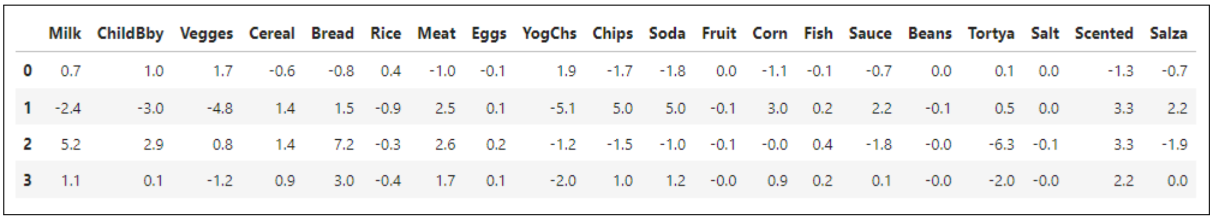


(a) Cluster Centers (2D)

(b) Cluster Centers (3D)

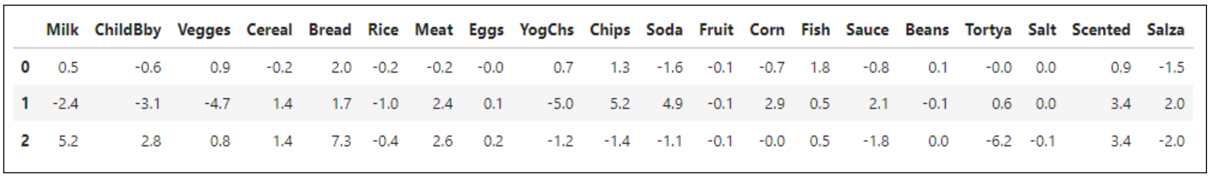
Figure 7: Cluster Centers (Visualized)

Question-9: What prototype amounts do you get back? What are the relative amounts? Are these completely realistic? Do you notice anything odd? Did anything become negative?



	Milk	ChildBby	Vegges	Cereal	Bread	Rice	Meat	Eggs	YogChs	Chips	Soda	Fruit	Corn	Fish	Sauce	Beans	Tortya	Salt	Scented	Salza
0	0.7	1.0	1.7	-0.6	-0.8	0.4	-1.0	-0.1	1.9	-1.7	-1.8	0.0	-1.1	-0.1	-0.7	0.0	0.1	0.0	-1.3	-0.7
1	-2.4	-3.0	-4.8	1.4	1.5	-0.9	2.5	0.1	-5.1	5.0	5.0	-0.1	3.0	0.2	2.2	-0.1	0.5	0.0	3.3	2.2
2	5.2	2.9	0.8	1.4	7.2	-0.3	2.6	0.2	-1.2	-1.5	-1.0	-0.1	-0.0	0.4	-1.8	-0.0	-6.3	-0.1	3.3	-1.9
3	1.1	0.1	-1.2	0.9	3.0	-0.4	1.7	0.1	-2.0	1.0	1.2	-0.0	0.9	0.2	0.1	-0.0	-2.0	-0.0	2.2	0.0

Figure 8: Reprojection Prototype (2D)



	Milk	ChildBby	Vegges	Cereal	Bread	Rice	Meat	Eggs	YogChs	Chips	Soda	Fruit	Corn	Fish	Sauce	Beans	Tortya	Salt	Scented	Salza
0	0.5	-0.6	0.9	-0.2	2.0	-0.2	-0.2	-0.0	0.7	1.3	-1.6	-0.1	-0.7	1.8	-0.8	0.1	-0.0	0.0	0.9	-1.5
1	-2.4	-3.1	-4.7	1.4	1.7	-1.0	2.4	0.1	-5.0	5.2	4.9	-0.1	2.9	0.5	2.1	-0.1	0.6	0.0	3.4	2.0
2	5.2	2.8	0.8	1.4	7.3	-0.4	2.6	0.2	-1.2	-1.4	-1.1	-0.1	-0.0	0.5	-1.8	0.0	-6.2	-0.1	3.4	-2.0

Figure 9: Reprojection Prototype (3D)

The prototype amounts don't necessarily makes sense to me, since as far as the last assignment was concerned, the prototypes were calculated by taking the center of mass for each of the clusters obtained through agglomerative hierarchical clustering. And since the values represented the number of tangible objects bought, they always had to be non-negative.

However, since the Principal Components are not necessarily tangible objects or hold any physical meaning and are only abstract entities defined as linear combinations of our original independent variables, they aren't bound by the same rules.

The data is no longer being operated upon in our original space, but rather in the more abstract PCA space. Thus gathering data and calculating cluster centers in PCA space, then reprojecting that abstract idea of a prototype to our regular space doesn't necessarily make absolute sense, especially since we discarded some of the attributes which originally existed in our agglomeration data.

> **Question-10:** If you projected the data onto all of the eigenvectors, why would this not help you with your data understanding? How many dimensions would you have?

⇒ The entire point of PCA is to obtain dimensionality reduction by utilizing the fact that not every attribute contributes the same usefulness towards the understanding of data. So, we use abstract entities called principal components which try to efficiently capture the explained variances of our independent variables / attributes. These principal componenets are abstract because they are essentially, in a way, linear combinations of our various attributes.

Now, these principal components do not hold equal importance, and we only require those that have maximal useful variance and we discard the rest.

If we use all of our eigenvectors, that would essentially be of no use, since we would first be performing a bunch of linear combinations, only to result in the same amount of understanding we had with our original data (if not less).

So since we would have 20 eigenvectors, it would have the same dimensionality as just using the original attributes.

Overall Conclusion:

It was interesting to observe how the percent explained variance versus principal componenet graph showed which principal componenets were of useful, and which were to be discarded by checking the drop in explained variance. Also that the cumulative sum of normalized eigenvalues looked essentially like a flipped percent of explained variance graph: their relation wasn't apparent to me initially, even though it feels obvious to see since its a running total.

I also ponder about how PCA is affected if there are a lot of highly correlated attributes in our data. The first eigenvector had a decent contribution from both Veggies and Soda. These two attributes are strongly correlated (negatively), as we saw in the last assignment.

And in general I was highly intrigued by how the transformation from the regular space to PCA space allowed for such drastic performance boost in terms of finding clusters and cluster centers. It would be an interesting exercise to try this with larger and larger test suites and compare the performance differentials. Also the simple fact that when I used two eigenvectors, I obtained 4 clusters, but using 3 yielded 3 clusters, which we know corresponds to our findings from the agglomerative clustering assignment.