HW_08:PCA and K-Means

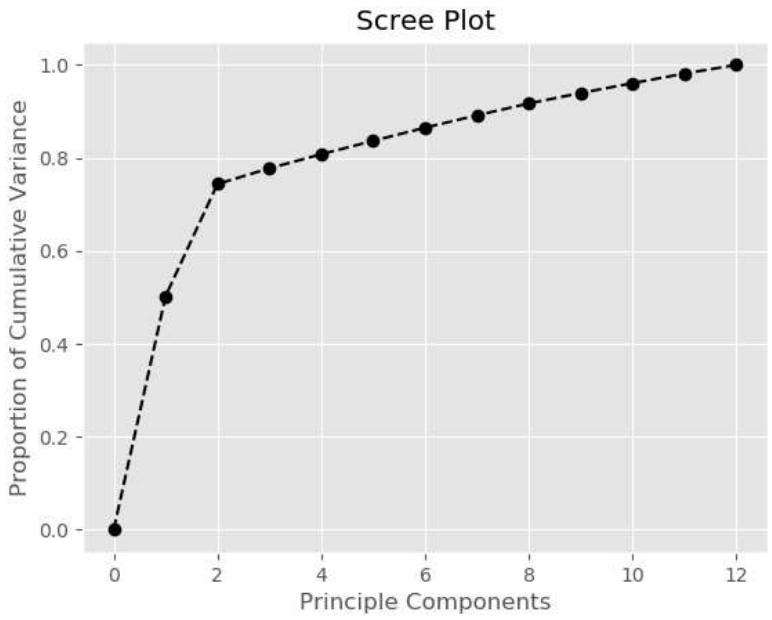
Q.1

By using 10 visits we are **marginalizing** the data. Marginalizing is important to account for any bias/uncontrollable error in a one or more particular visits. It basically makes up for the erroneous measurements which are naturally present that we can't get rid of due to the nuisance variable. In this way, Marginalization makes the data points **commensurate**.

Q.5

The plot of cumulative normalized eigen values against the each of the eigen values where the cumulative normalized eigen value is the total variation as explained by each of the eigen value.

This plot is known as the **scree plot (for cumulative variance).**



The x-axis has the principle components (eigen values) in decreasing order of the proportion of cumulative variance explained.

The Y-axis is the cumulative variance for each of the eigen value respectively.

Q.6

The first two eigen vectors are:

Eigen_Vector_1

[-0.06751563 -0.14562291 0.2854674 -0.27575588 -0.17465001 0.44101104

-0.01397 0.0013735 0.38308582 -0.28350631 -0.52146053 0.30407335]

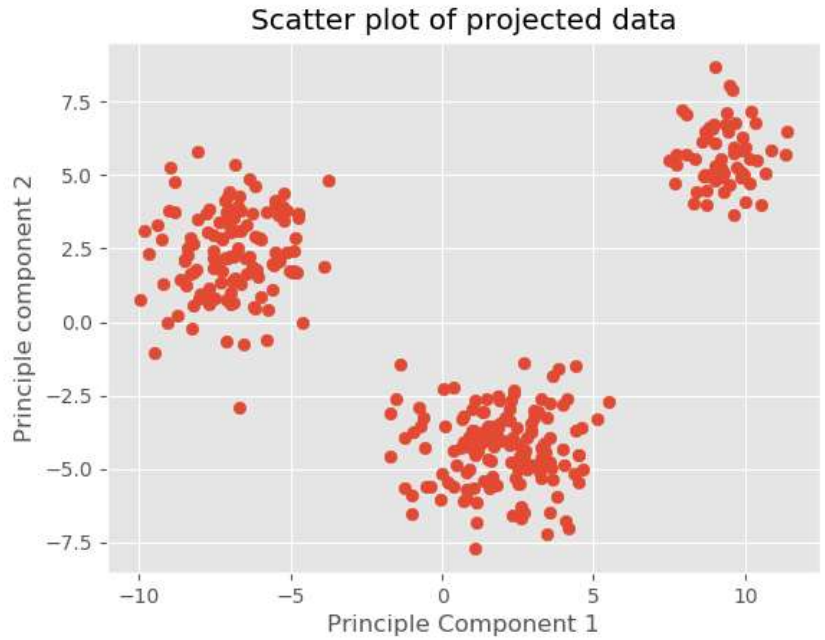
Eigen_Vector_2

[-0.50936981 -0.19195449 -0.06421965 -0.50711915 -0.37880847 -0.26258848

0.39216349 0.0026962 0.00910849 0.16036328 0.21096546 0.07647266]

Q.7

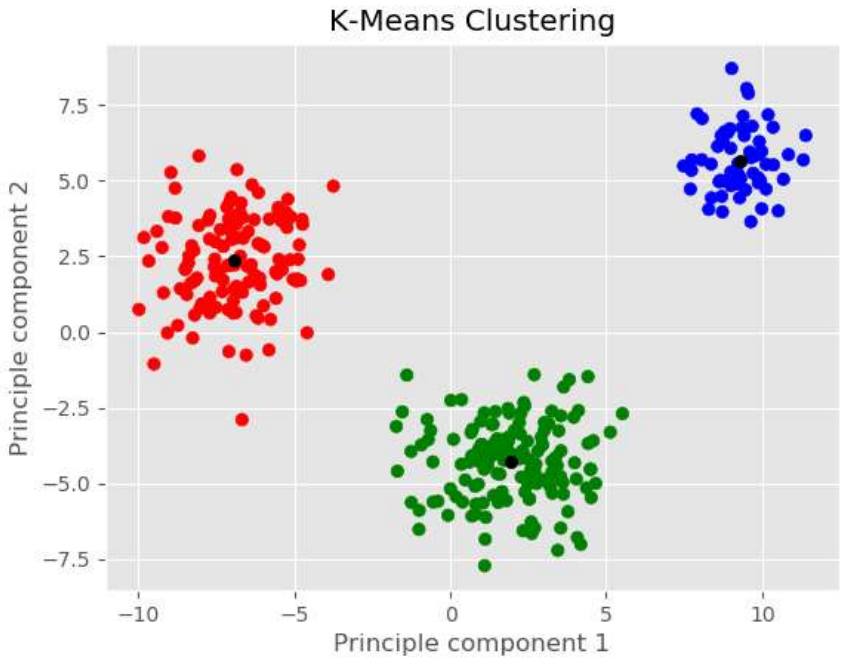
The 2-D scatter plot of the projected points is as follows:



Q.8

From the plot in Q.5, we can clearly see that after the third eigen value, the variation is consistent. Thus, we take k=3 in our problem.

The package (sklearn) uses the Eucliean distance by default to perform K-means clustering.



In this plot we can clearly see the 3 clusters.

The black dots in each cluster are the centroids of each of the clusters respectively.

Q.9

The centroids of the 3 clusters are respectively,

These centroid tell us which eigen vector has more effect on the cluster comparatively,

Cluster_1: [9.28615107 5.66363139]

Cluster_2: [1.94444433 -4.30168234]

Cluster_3: [-6.93926413 2.35285764]

Q.10

The prototypes of the three clusters with respective centroids are as follows:

A. cluster centers for centroid [1.94444433 -4.30168234]

Milk	7.683012
PetFood	4.788862
Veggies	7.700764
Cereal	7.998389
Bread	6.518403
Rice	8.251189
Meat	3.914953
Eggs	5.047452
Yogurt	5.322918
Chips	2.931013
Cola	2.876762
Fruit	5.078316

B. cluster centers for centroid [-6.93926413 2.35285764]

Milk 4.893180

PetFood 4.805165

Veggies 4.737403

Cereal 7.073479

Bread 5.549147

Rice 2.585970

Meat 6.648726

Eggs 5.053193

Yogurt 1.980308

Chips 6.516744

Cola 8.913144

Fruit 2.885908

C. cluster centers for centroid [9.28615107 5.66363139]

Milk 2.111302

PetFood 1.806855

Veggies 9.156613

Cereal 0.920269

Bread 1.461229

Rice 8.872186

Meat 7.720422

Eggs 5.084405

Yogurt 8.226191

Chips 2.447663

Cola 1.150689

Fruit 8.072808

From the three cluster prototypes we can see that,

Cluster 1 has relatively higher average values for all attributes in comparison to the other two clusters.

In comparison to other clusters, Cluster 2 has moderate average values for all attributes(prototype) and higher average value (prototype amounts) for the variable like Cola and chips.

In comparison to other clusters, Cluster 3 has very low average prototype amounts for the attributes like cereal, milk, bread and high values for attributes veggies, rice ,yogurt, fruits.

From the previous homework(Agglomerative Clustering), the above prototype values clearly convey the three clusters : Family member, party animals and gluten free clusters respectively.

These values are completely realistic and I did not find anything odd.

Q.11

What I did:
I used the pandas library to find the covariance matrix and then the numpy library to find the eigen values and corresponding eigen vectors. I stored the transpose of the eigen vector for further calculations. I further normalized the eigen value and found the cumulative sum of the normalized eigen values which is equivalent to the total cumulative variance explained by each of the eigen value. I then projected the first two eigen vectors (which explained about 70% of variation) on the centered data to get the values of the two principle components on our data (which are 337 values for each of the two vectors).
Based on this new projected principle components data I performed K-Means clustering using sklearn package. I extracted the labels i.e cluster class for each row(customer) and the centroids for each of the cluster (which are two co-ordinates, based on 2-D projected data).

I then projected the two eigen vectors on the centroids of the clusters, added the two values obtained by projection (both eigen vectors) and added the respective means of the attributes to it.

What I learned:

Importance of Eigen values and Eigen vectors: A vector whose direction remains unchanged by a linear transformation that has a wide application in face recognition.

Importance of Marginalization as explained in Q1.

Importance of PCA: I learned the importance of PCA in dimension reduction. In this problem, to perform K-means clustering. Also learned how PCA can be used for feature selection by comparing the eigen vectors.

Scree-plot to obtain K.

Use of packages like sklearn for K-means that simplifies a lot of work.