Rohit Ravishankar
rr9105@rit.edu

Home Work 7

CSCI - 720 Big Data Analytics

1.

You are provided with the file `HW_AG_SHOPPING_CART_v???.csv`. It contains data for the number of times various categories of items (attributes) were purchased by guests, for 10 different visits.

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

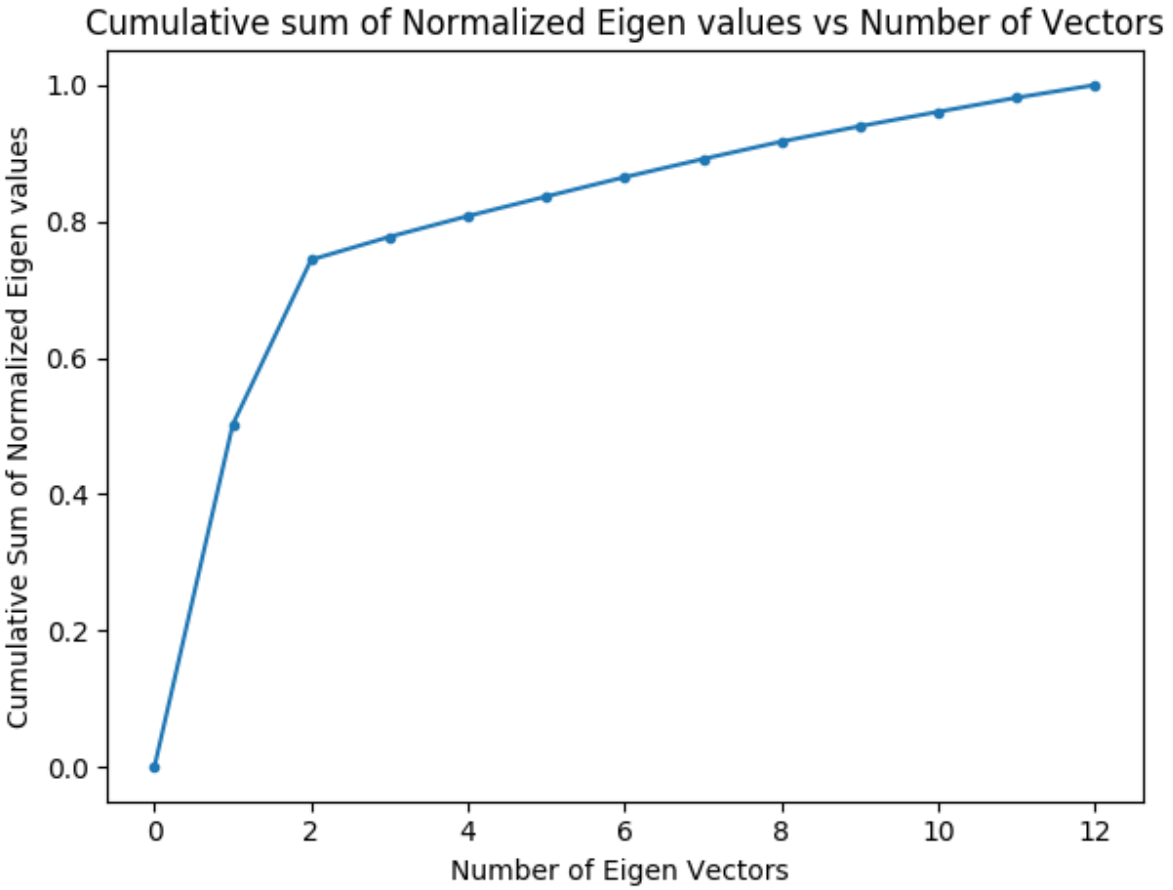
Ans.

It is also called marginalized data because collecting larger amount of data is a form of data cleaning or noise removal

5.

Normalize eigenvalues by dividing each by the total of all the absolute values, and plot the cumulative sum of these normalized eigenvalues. The plot will start at the origin for no eigenvalues, and end with a y value of 1.0 for all of the eigenvalues. **Plot: Report the plot in your final report.**

Ans.



6.

Print out the first two eigenvectors – the eigenvectors associated with the eigenvalues that have the largest two eigenvalues. Look at the components of each one.

Q: Why does this tell you about the attributes?

Ans.

```
highest vector 1: ['-0.068', '-0.146', '0.285', '-0.276', '-0.175', '0.441', '-0.014', '0.001', '0.383', '-0.284', '-0.521', '0.304']
highest vector 2: ['-0.509', '-0.192', '-0.064', '-0.507', '-0.379', '-0.263', '0.392', '0.003', '0.009', '0.160', '0.211', '0.076']
```

What is important is to note is:-

- the absolute value of the components and
- the relative directions with respect to each other.

From the above 2 Eigen vector values, we see that at index position 7, the value is close to 0. All the other attributes have significantly larger values and hence, the value at position

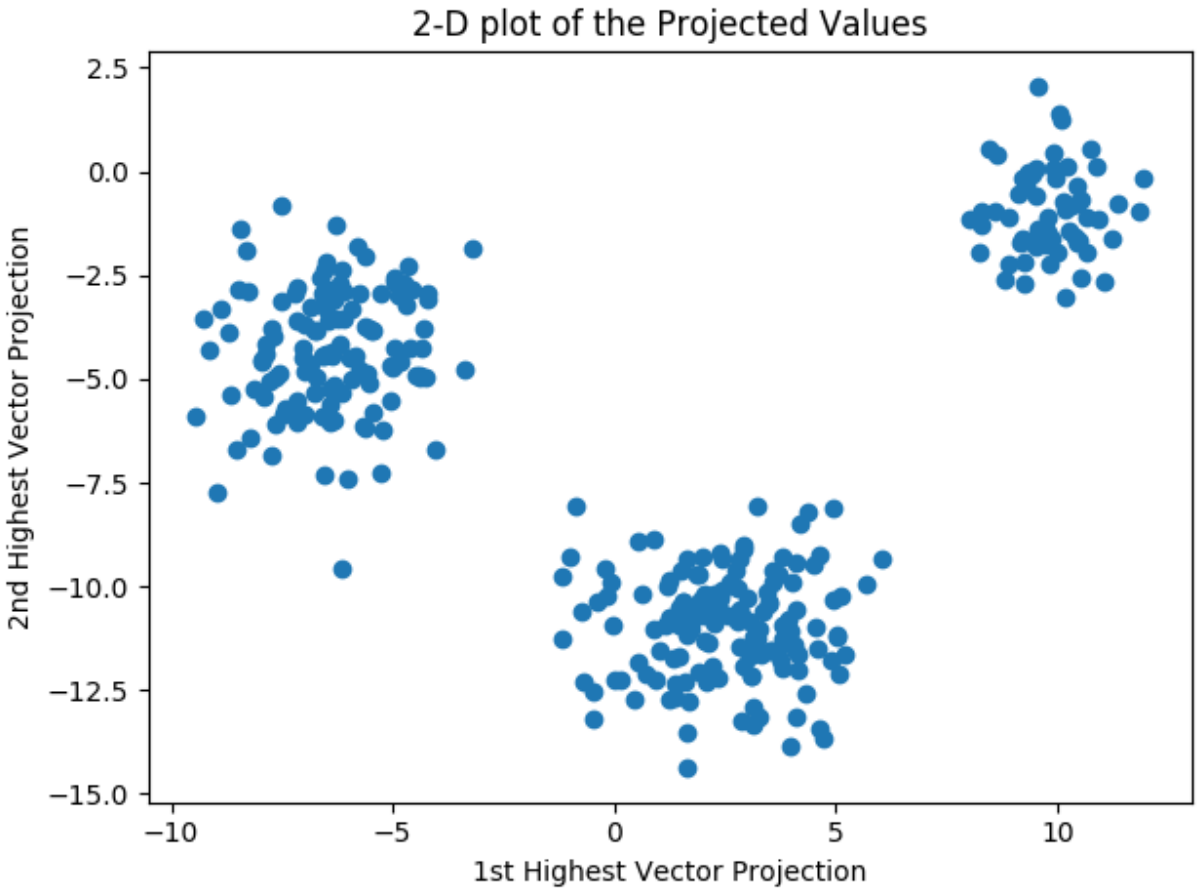
7 can be ignored. The corresponds to eggs on the data set which is akin to what we saw the previous homework, i.e., it can be ignored.

Again, if we look at index position 5, 6 & 10, we see that the values for the items are opposite to each other indicating that one individual is consuming significantly more or less as compared to the other individual.

7.

Project the original Agglomeration data onto these first two eigenvectors.
PLOT: Generate a 2D plot of these projected points, and show a scatter gram of this 2D plot of points.

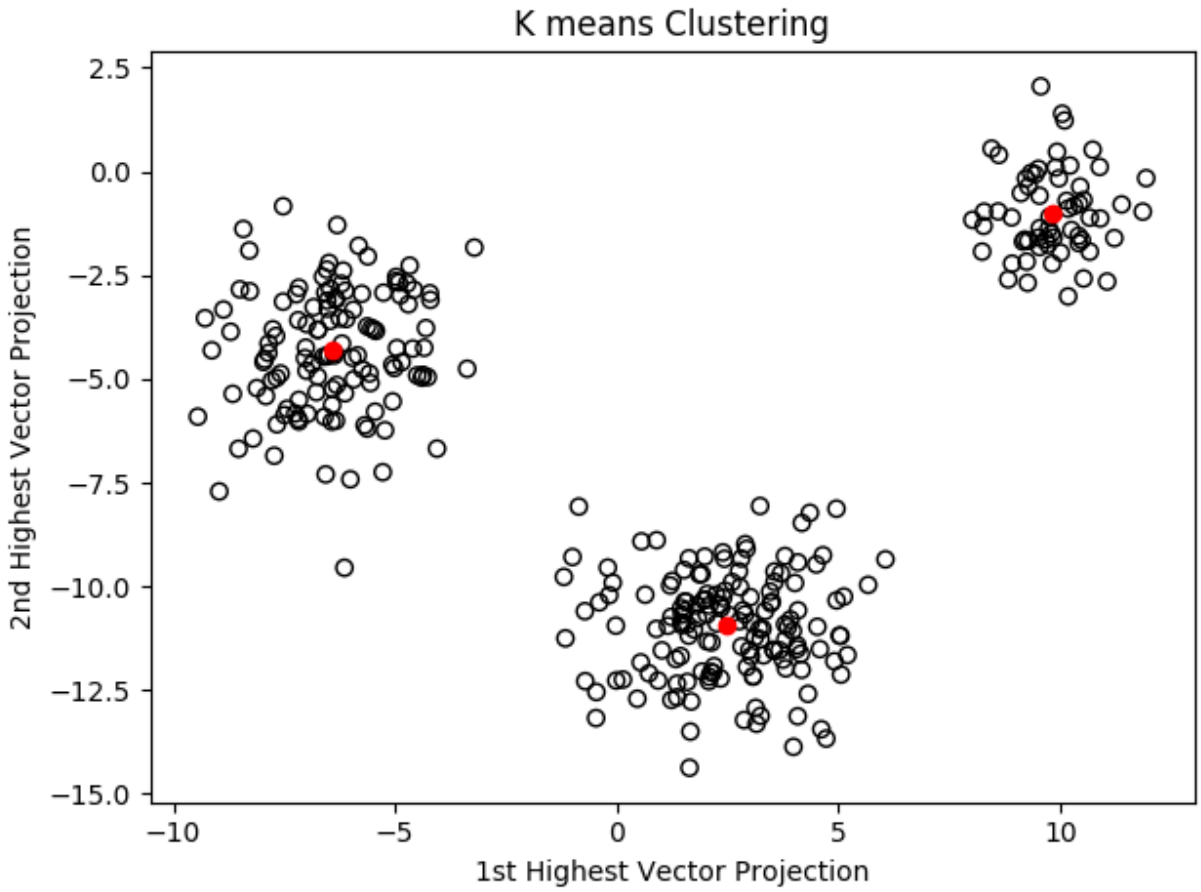
Ans.



8.

In this new, projected, two dimensional space, perform k-Means clustering using a package of your choice. For simplicity, use the Euclidean distance. Use the k value that you would expect to get from the number of clusters you can see in your plot from part 5.

Ans.



9.

Find the center of mass (the average values) of each of the k clusters that you got out of k-Means. These are 2D vectors. **Print all k of these 2D vectors. What do they tell you?**

Ans.

```
cluster center 1: (2.482 ,-10.957)
cluster center 2: (-6.402 ,-4.302)
cluster center 3: (9.823 ,-0.991)
```

The values tell you the centroids of clusters of data that are created. They can be used to classify new data. In PCA, the cluster centers are the nearest approximation to the mean value of the data. They tell us how well separated are the clusters.

10.

Multiply these centers of mass back times the first two eigenvectors.

What prototype amounts do you get back? What are the relative amounts? Are these completely realistic? Do you notice anything odd?

Ans.

```
Vector 1: ['11.037', '7.365', '7.035', '10.495', '9.340', '9.595', '1.292', '5.597', '6.474', '3.163', '2.018', '5.540']
Vector 2: ['6.870', '6.004', '2.695', '8.193', '6.994', '2.553', '2.649', '4.226', '1.755', '5.371', '6.677', '1.971']
Vector 3: ['6.711', '5.629', '9.737', '4.663', '5.529', '11.462', '6.343', '6.880', '10.624', '3.926', '1.538', '9.781']
```

Yes, the relative amounts are realistic. No, I did not notice anything odd.

Another interesting point to note is that if we add overall average values to the to the reprojected values, the re-projection values aren't very different from the actual values.

11.

Write up an overall summary of what you did, what you learned, and what you found from this experiment.

Ans.

We learnt how to implement PCA to perform dimensional reduction and feature selection. I read the file into a pandas data frame and generated the covariance matrix. Using, the covariance matrix I computed the Eigen vectors and Eigen values. Subsequently, I sorted the Eigen values in terms of the highest to lowest absolute value and normalized them to plot the cumulative sum of normalized Eigen values. Based on the values obtained in Eigen vectors, the largest values can be used determine the most important features and the smallest values to determine the least important features. Hence, we remove *'eggs'* from consideration as it has a very small value in comparison to the rest of the features. Further, I used *sklearn* to compute the K-Means for the given dataset. The value of k in K-Means was 3 because, the scatter plot obtained had 3 distinct clusters. For some of the prototype amounts in the last step some of the values were negative & unrealistic.