Nutrient analysis of pizzas

The data set contains measurements that capture the kind of things that make a pizza tasty.

Question 1: understanding the customer preferences

Question 2: identify specific groups(clusters) of pizzas that have similar nutritional content

Question 3: Are there any correlations between certain nutritional factors, such as the amount of protein and fat in the pizza?

Question 4: any important factors that be derived from the variables and checking their relationships

brand -- Pizza brand (class label)
id -- Sample analysed
mois -- Amount of water per 100 grams in the sample
prot -- Amount of protein per 100 grams in the sample
fat -- Amount of fat per 100 grams in the sample
ash -- Amount of ash per 100 grams in the sample
sodium -- Amount of sodium per 100 grams in the sample
carb -- Amount of carbohydrates per 100 grams in the sample
cal -- Amount of calories per 100 grams in the sample

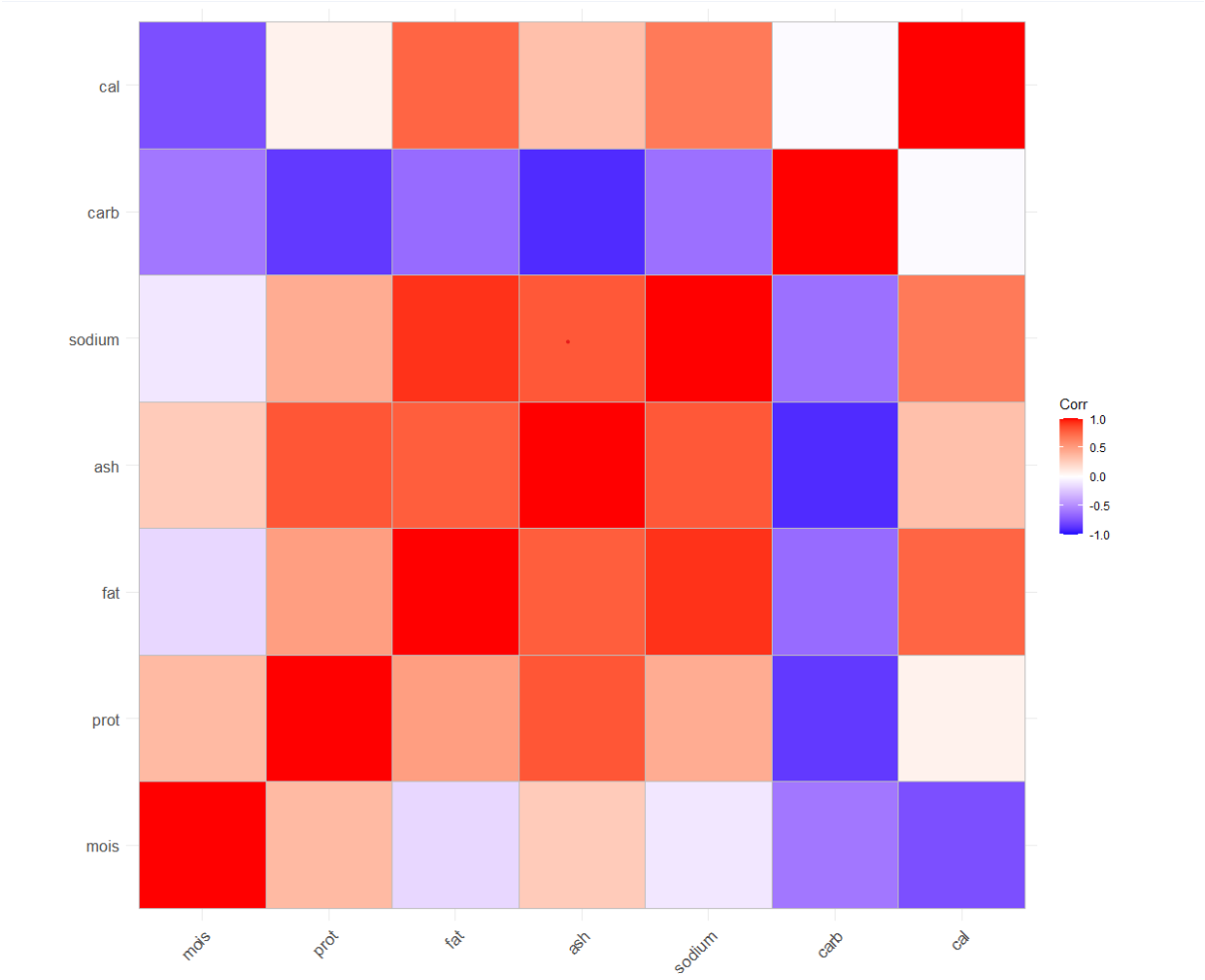
link: https://data.world/sdhilip/pizza-datasets

The dataset contains 300 observations and 9 variables. To perform the analysis, the dataset must contain only numerical data. The dataset includes an “id” column, and the “brand” column, which categorizes the pizzas into six different types. Since the “id” column does not contain any meaningful information about the dataset and is only used to differentiate between observations, it needs to be removed before performing the analysis. Also, the “brand” column, being a categorical variable, does not add any numerical information to the analysis and does not contribute to the understanding of the variation in the data. Therefore, it is not relevant to the analysis and needs to be removed as well.

Data Scaling-

For analysis, the data must be standardized to ensure that all variables are on the same scale and carry equal weight in the analysis. This involves transforming each variable to have a mean of zero and a standard deviation of one.

Correlation Matrix:



Carbohydrates is negatively correlated with almost all other variables and not correlated with calories.

Calories, Proteins, and Carbohydrates are not correlated as all 3 are white in color i.e, 0.0 value in correlation matrix.

PCA

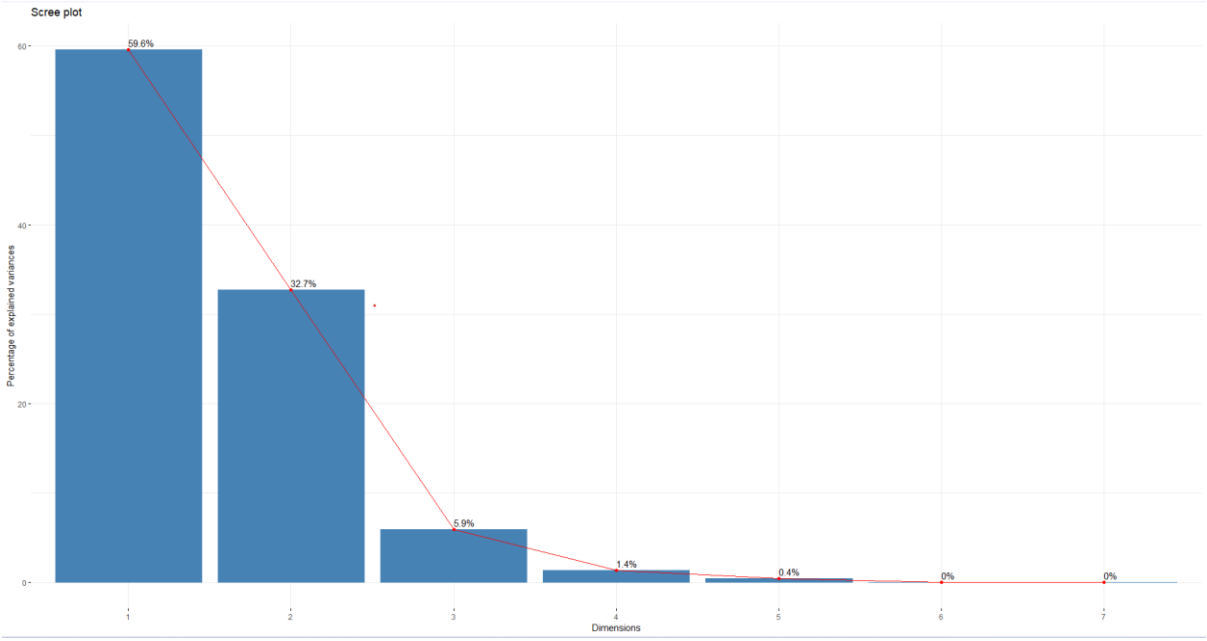
```
pca_data$rotation[,1:2]
```

```
##                PC1        PC2
## mois     0.06470937 -0.6282759
## prot     0.37876090 -0.2697067
## fat      0.44666592  0.2343791
## ash      0.47188953 -0.1109904
## sodium   0.43570289  0.2016617
## carb    -0.42491371  0.3203121
## cal      0.24448730  0.5674576
```

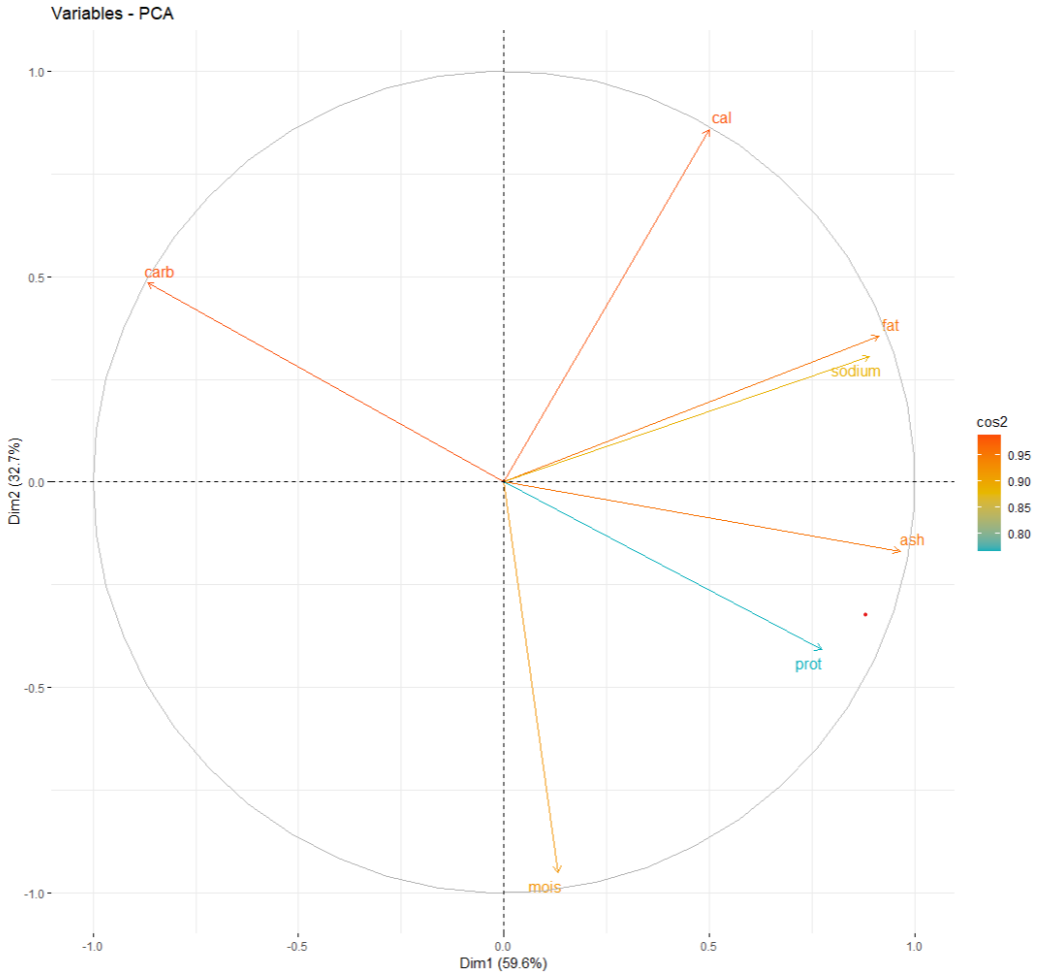
```
eigenvalues <- summary(pca_data)$importance[2,]
eigenvalues
```

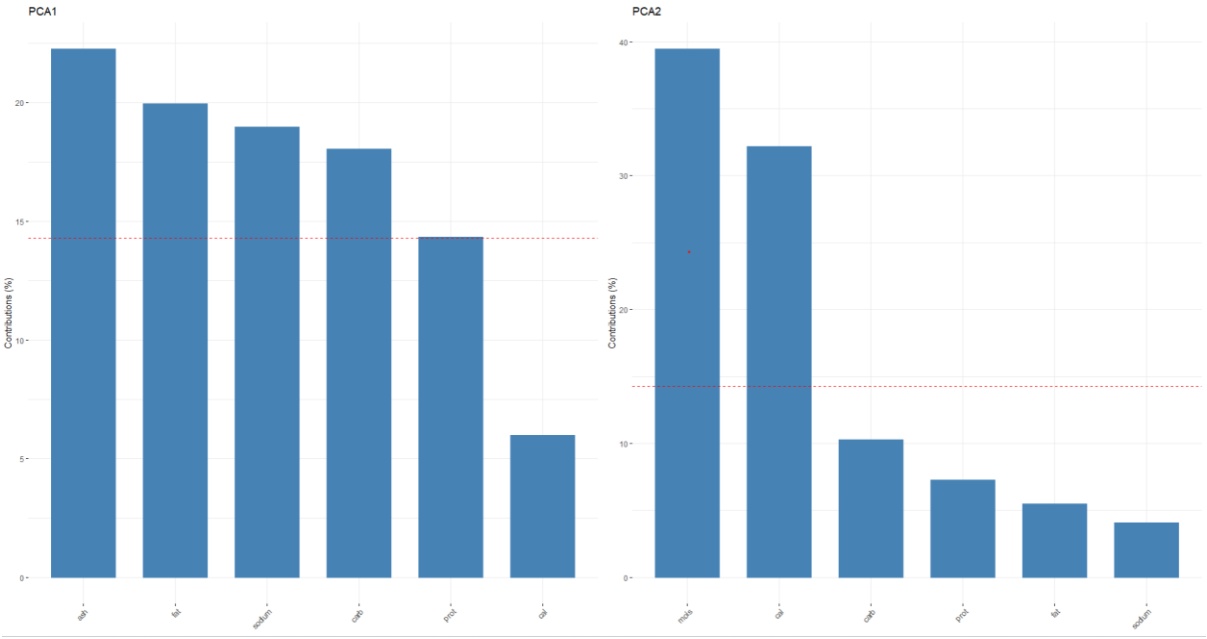
```
##     PC1     PC2     PC3     PC4     PC5     PC6     PC7
## 0.59597 0.32721 0.05922 0.01360 0.00395 0.00005 0.00000
```

The results of the loading matrix show that the majority of the variation in the dataset is captured by the first principal component (PC1), which is primarily driven by the levels of fat, ash, sodium, and calories in the pizza. The second principal component (PC2) explains a smaller proportion of the variance, and is primarily driven by the levels of mois, prot, and carb in the pizza.



The first two components, PC1 and PC2, have the most significant information with the highest eigenvalues. These two principal components explain almost 92.3% of the total variation in the data, indicating their importance. Hence, we can conclude that PC1 and PC2 capture the essence of the dataset effectively, while the rest of the components are comparatively less important in explaining the variation in the data.



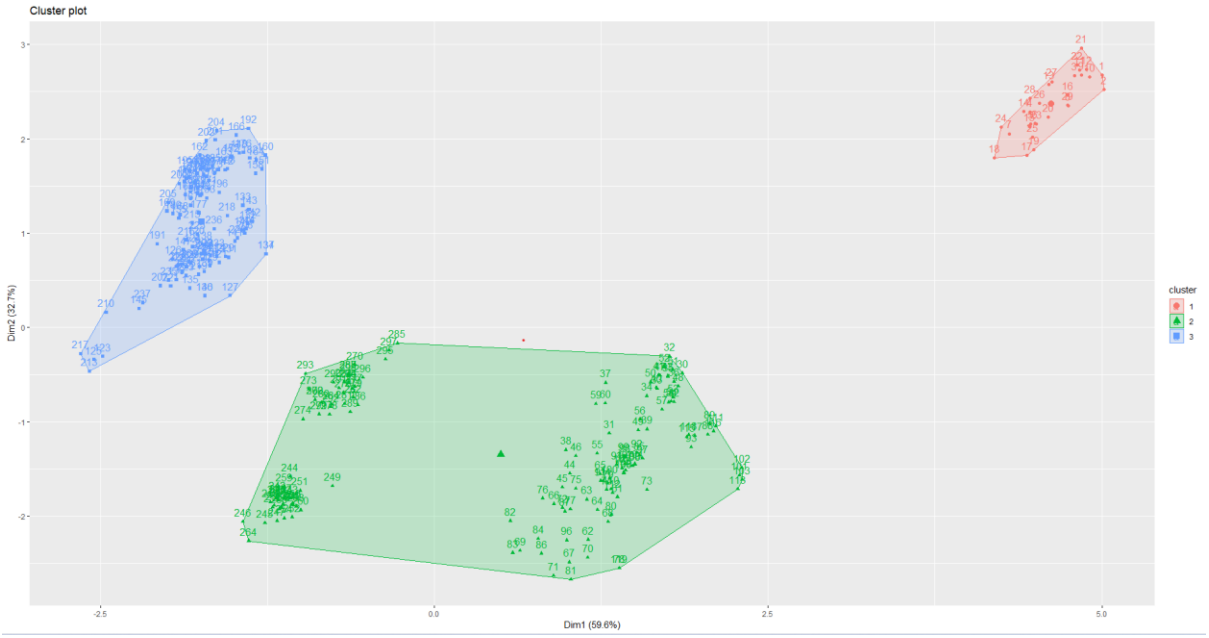


Five variables make a significant contribution to the first principal component, whereas the second principal component mainly comes from only two variables, accounting for over 70% of its variation.

Kmeans Clustering:

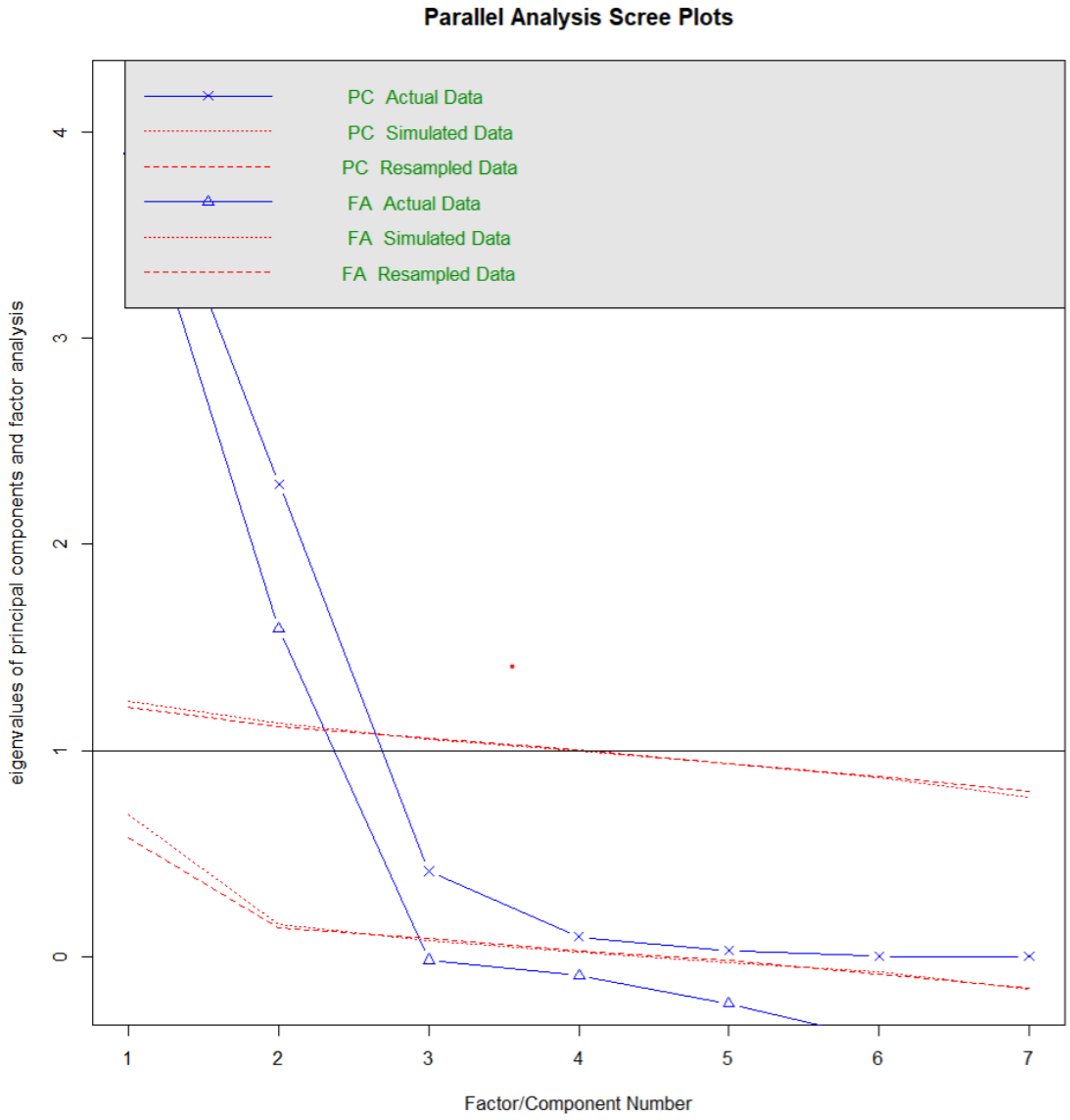
```
## *** : The D index is a graphical method of determining the number of clusters.
##                 In the plot of D index, we seek a significant knee (the significant peak in Dindex
##                 second differences plot) that corresponds to a significant increase of the value of
##                 the measure.
##
## *******************************************************************
## * Among all indices:
## * 3 proposed 2 as the best number of clusters
## * 12 proposed 3 as the best number of clusters
## * 1 proposed 5 as the best number of clusters
## * 7 proposed 8 as the best number of clusters
##
##                     ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  3
##
##
## *******************************************************************
```

determined the optimal number of clusters (k) in a k-means clustering algorithm, which is 3.

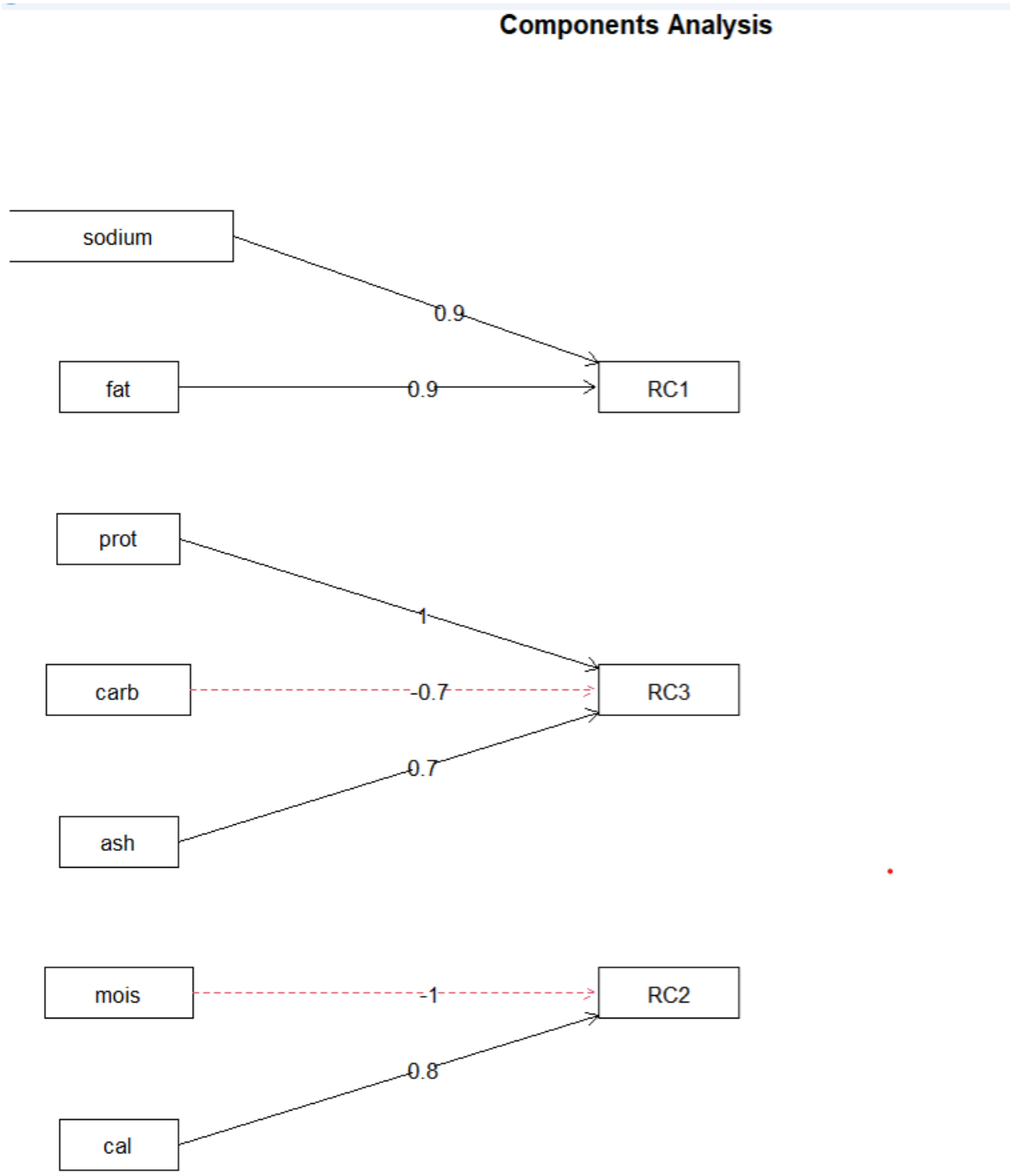


The centroid of each cluster is represented by a larger point with a white border.

Exploratory Factor Analysis:



Parallel analysis suggests that the number of factors = 2 and the number of components = 2



In RC1 factor, sodium and fat contributed positively with 0.9

In RC3 factor, protein and ash contributed positively whereas carbohydrates impact negatively

In RC2 factor, comprises of calories and moisture negatively impact the factor

Conclusion:

PC1 is an index of energy supply (Protein(2) + Ash(4) + Carbohydrates(6))

PC2 is an index of the heaviness of pizza (Calories(7))

These findings suggest that the pizza dataset can be summarized by two main factors: the level of calories, fat, ash, and sodium, and the level of mois, prot, and carb. These factors can be useful for understanding consumer preferences and developing new pizza products.

K-mean clustering was performed with 3 clusters on the pizza_data dataset

EFA suggests the number of factors that can be derived is 2.