Partitioning the Parts of a Pizza

Amar Lakshya (Student ID: 2096845)

School of Computer Science, University of Birmingham

Abstract

Pizza is one of the most popular fast food in the world. In this report, experiments were conducted over a sample pizza dataset to develop a better understanding of the various attributes of pizza and their relationships. A number of data-analysis techniques like Principal Component Analysis and co-ordinate projections were employed to understand the characteristics of 10 anonymized brand names in the dataset and 7 different attributes of each pizza.

Keywords: Food, Pizza, PCA

1. Acknowledgement

I would like to thank Prof. Martin Russell for this educational and challenging assignment, and Dhilip Subramanian for his pizza dataset.

2. Introduction

The story of making food in a Pizza like fashion goes back to the Neolithic Age [1]. There have been several notable instances of Pizza consumption in almost all kinds of cultures ranging from the Persian to Greeks [2] [3] This is the introduction. Although, it perfectly clearly that its consumption has only increased in the recent times and topping choices have been extensively discussed, a lot of attributes of the food still remains hidden in mystery.

Through this project, we aim to explore those attributes and their hidden relationships with each other that might influence qualitative aspects of a Pizza.

3. Methods

The results were obtained by using the following python tools (libraries): *numpy, pandas, matplotlib, sklearn.* [4]

The entire source code of the project containing the generated images and the analysis can be found at: Github.

4. Dataset

The sample dataset has been retrieved from the following url on the Data World website. The dataset is in the ***csv*** file format and is **","** separated It is comprised of the following columns:

1. **brand** – Pizza brand (class label)
2. **id** – Sample analysed
3. **mois** – Amount of water per 100 grams in the sample
4. **prot** – Amount of protein per 100 grams in the sample
5. **fat** – Amount of fat per 100 grams in the sample
6. **ash** – Amount of ash per 100 grams in the sample
7. **sodium** – Amount of sodium per 100 grams in the sample
8. **carb** – Amount of carbohydrates per 100 grams in the sample
9. **cal** – Amount of calories per 100 grams in the sample

The following things are to be noted about the dataset:

- The ***id*** attribute has been removed from the data-set since it doesn't especially contribute to the analysis of the data.
- The ***brand*** attribute is the class label that already comes with the dataset and thus can be used initially.

5. Preprocessing

In order to investigate the data any further it was necessary to first check whether the entire dataset had standardised values. To do this, the scale of the measurements of the data had to be checked. This was done by creating a box plot which can be seen in Figure 1.

As is seen from the figure, all the attributes contribute different units to the data and therefore it was necessary to standardize the data.

Standardisation for this dataset was done using the *preprocessing* class from the *sklearn* library [5] and applying the *fitTransform* function. After applying the standardisation, a new box plot was produced and can be seen in Figure 2.

6. Labelling

Since the original dataset came labelled with the class field *brand*, there is no need for a labelling scheme and the values in the *set* of *brand* namely: *A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J'* can be used.

7. Principal Component Analysis

The standardised data was then put through the *pca* function of the *sklearn* library [6] to generate the eigenvectors, eigenvalues, and the coefficients of the eigenvalues. To find the most useful eigenvectors from the PCA step, a scree plot of the percentage of variance was plotted and can be seen in Figure 3.

As can be seen from the scree plot, the first two principal components namely: *PC1, PC2* contribute nearly **59%** and **32%** of variances respectively. Therefore, only these two principal components were chosen for further analysis.

A 2D projection was then generated to cover all the classes through the two chosen principal components and can be seen in Figure 4. This can be considered a good result as cluster of different classes can be identified uniquely. It might be improved further by adding another principal component but since the percentage of variance was too low, such a decision was not considered wise.

A biplot was also generated to show both the coefficients of the eigenvalues for each attribute and the principal component score for each data point and can be seen in Figure 5. Furthermore, Table 1 and Table 2 show the corresponding eigenvectors and eigenvalues.

The direction and the length of vectors produced indicates the contribution to the principal components from each attribute. For example it can be observed from Fig 5 that attributes such as *mois, cal, prot, fat, ash, sodium* being on the the right side of the y-axis most contributed to *PC1*, while *carb, fat* being above the x-axis most contributed to *PC2*.

8. Coordinate Projections

As vectors close to each are bound to have strong correlations, A 2D projection of the closest pair namely, *fat* and *sodium* was generated and can be seen in Fig 6. It did look promising and hence 2D projections for all combination of the attributes were generated to be studied. (They can be found at Github). Consequently, *mois* was found to be the most important attribute in all the 2D projections as can be seen from Fig 7.

It worth stating that as clear from the biplot, the *mois* vector makes a relatively small angle with *prot* and *ash* vectors and they seem to be positively correlated. This can also bee verified by looking at the 2D projections for *prot* and *ash* (especially with *mois*) in which both act as good attributes to cluster the different classes neatly.

9. Histogram Analysis

Histograms for each attribute frequency for all the classes were generated and analysed (They can be found at Github). Consequently, a few observations were made:

- The histogram for *mois* attribute shown in Fig 8 , clearly demonstrates that this attribute captures all the class values effectively
- The histogram in Fig 8 for *fat, sodium* and *cal* clearly shows segmented large values for the *brand A*. These three attributes can also be seen contributing the most to the PC1 principal component as shown from the Fig 5 as welll as Table 1.
- The histograms in Fig 8 also shows that *brand* classes *G, H, I* have the most negative values in the *fat, sodium, cal* attributes.
- The historgrams also show that the brand classes *G, H, I* have positive values for only the *carb* attribute. Furthermore, From Fig 5, and Table 1 , we can see that *carb* contributes the most to PC2.

10. Conclusion

Standardisation of data was necessary and all calculations were based on the standardised dataset. The following were the observations noted at the conclusion:

- Adding a 3rd principal component can increase the accuracy of predictions.
- Projections of *mois* seems to be the best method to predict different brands of pizzas in the dataset.
- There seems to be close relationships between *fat, sodium, cal* attributes as derived from PCA.
- Brand *A* seems to be easy to predict using PC1 and usually contains high amounts of *fat, sodium* and *cal*.
- Brand *G, H, I* seem to be predicted by using PC2 and are usually low in *fat, sodium* and *cal* values.
- *mois, ash, prot* seem to be a good set of predictors for classes.
- Further clustering techniques can definitely be applied to the dataset to procure more knowledge based on the 2D projections produced.

References

[1] Perry, C. *A Stone-Age Snack : History: Pizza topped with tomatoes, pepperoni and cheese is only 100 years old.* Los Angeles Times,1991

[2] Barrett, L *Pizza, A Slice of American History*, 2014,13.

[3] Buonassisi, R. *Pizza: From its Italian Origins to the Modern Table. Firefly.* , 2000, 78.

[4] NymPy, Pandas,Matplotlib, Scikit-learn

[5] Importance of Scaling

[6] Sci-kit learn PCA function docs

PCs	mois	prot	fat	ash	sodium	carb	cal
PC1	0.064709	0.378761	0.446666	0.471890	0.435703	-0.424914	0.244487
PC2	-0.628276	-0.269707	0.234379	-0.110990	0.201662	0.320312	0.567458
PC3	-0.421669	0.746027	-0.199309	0.056273	-0.455169	0.052237	0.113316
PC4	-0.220722	-0.010593	-0.507042	0.552399	0.446277	0.334339	-0.279263
PC5	-0.006470	-0.387983	0.173368	0.670886	-0.602614	0.007437	0.078003
PC6	0.446450	-0.000172	-0.525403	0.058861	0.003131	-0.000509	0.721914
PC7	0.418569	0.276765	0.377672	0.056021	-0.000524	0.776068	0.012060

Table 1: Table showing PCA eigenvectors for each attribute

PCs	brand
PC1	4.185734
PC2	2.298118
PC3	0.415949
PC4	0.095493
PC5	0.027770
PC6	0.000339
PC7	0.000010

Table 2: Table showing PCA eigenvalues over the class

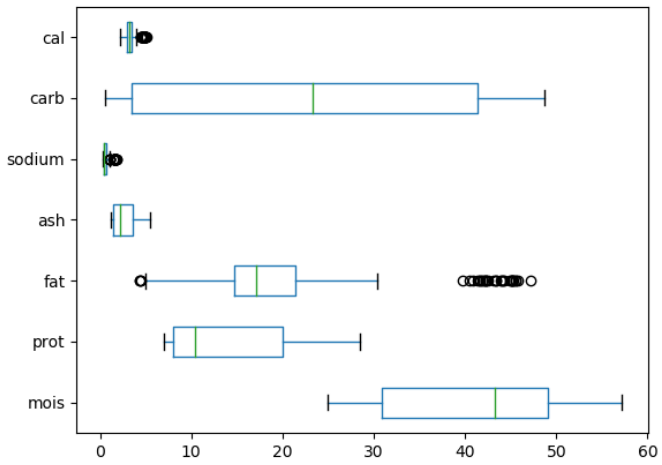

Figure 1: A box plot showing the raw ranges of the dataset

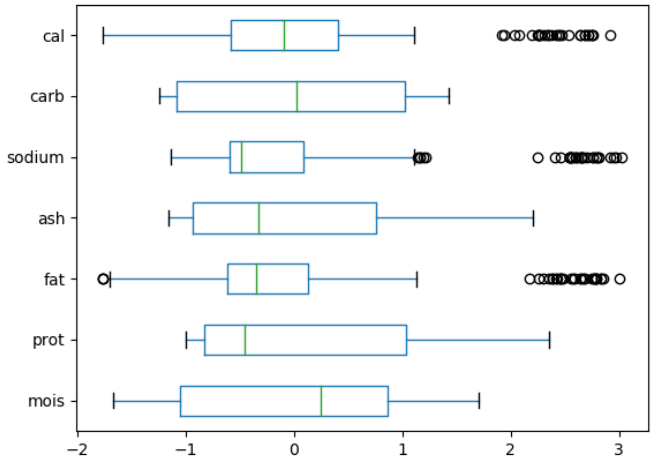

Figure 2: A box plot showing the standardised values of the dataset

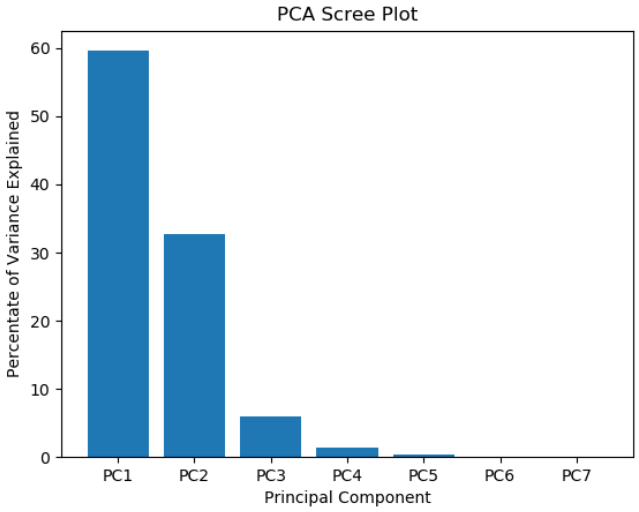

Figure 3: A scree plot over the various principal components

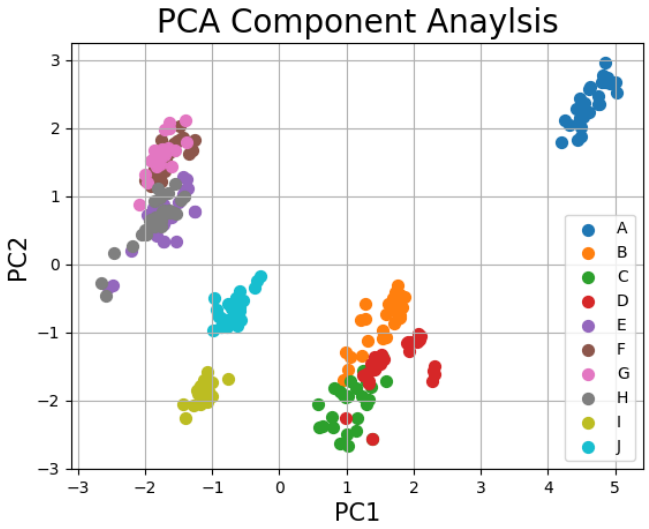

Figure 4: A 2D Projection over the two Principal Components

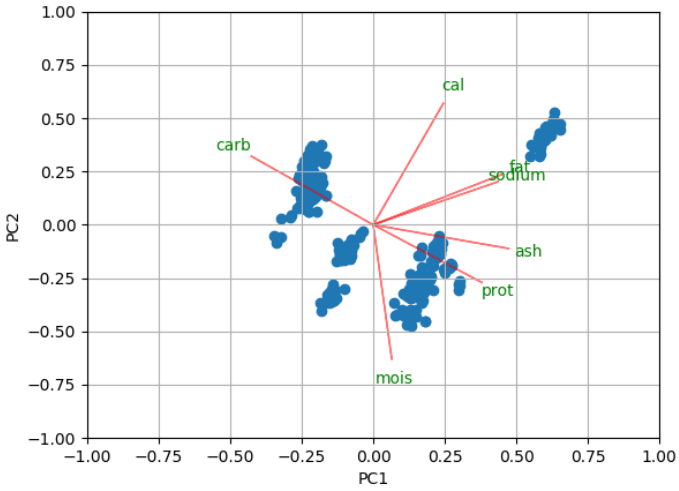

Figure 5: A biplot presenting the coefficients for each attribute

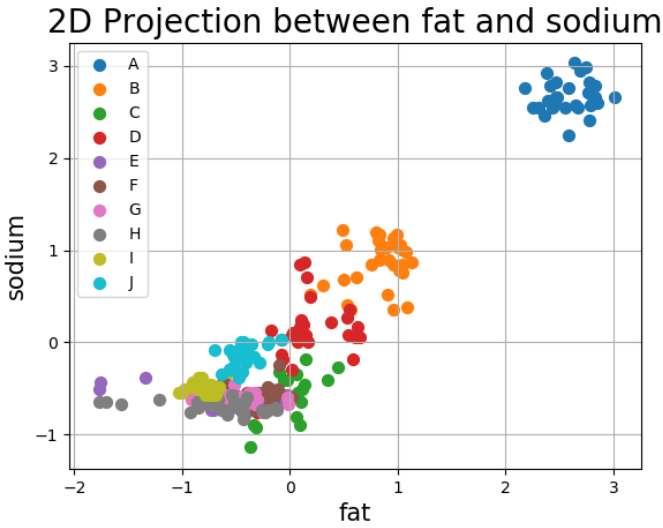

Figure 6: A 2D Projection over Fat and Sodium

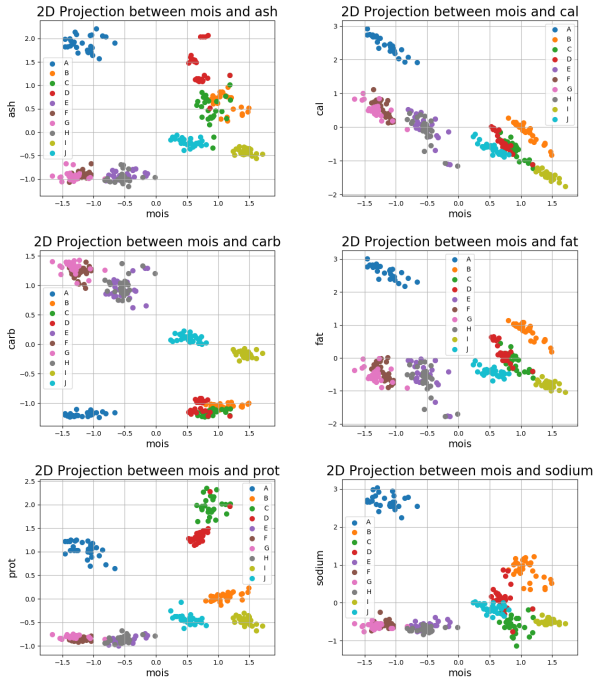

Figure 7: 2D Projections of *mois* with other attributes

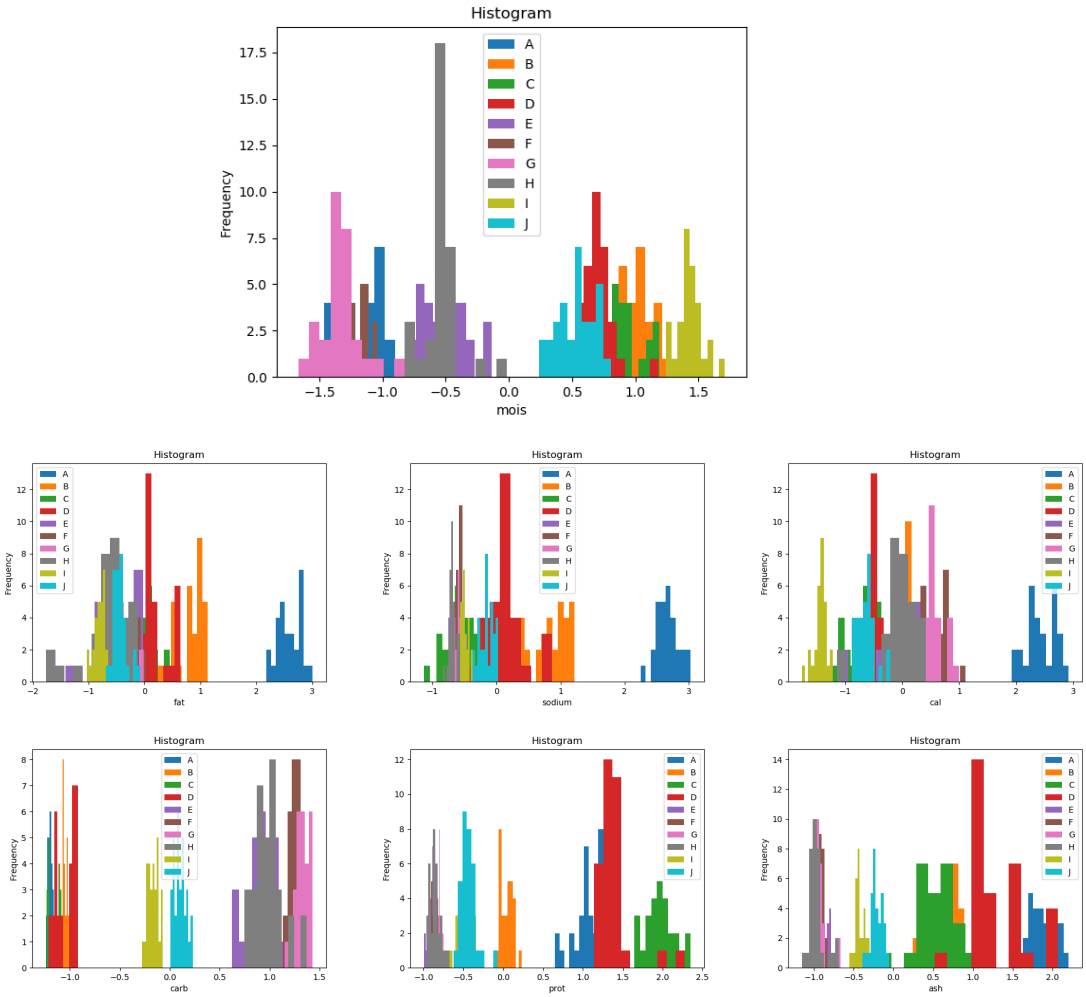

Figure 8: Histograms of *mois, fat, sodium, cal, carb, prot, ash*