PRINCIPAL COMPONENT ANALYSIS

HARSHAL SHIVAJI HOLAM
2102451

What is PCA ? :-

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analyzing data much easier

So to sum up, the idea of PCA is simple — reduce the number of variables of a data set, while preserving as much information as possible.

Dataset:-

The dataset was downloaded from this link: https://data.world/sdhilip/pizza-datasets and it contains information on the micronutrients of various pizzas.

About variables:-

The variables in the data set are:

- brand - Pizza brand (class label).
- id - Sample analised.
- 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.

Analysis:-

Analysis through MINITAB

OUTPUT

Eigenanalysis of the Correlation Matrix

Eigenvalue	4.1718	2.2905	0.4146	0.0952	0.0277	0.0003	0.0000
Proportion	0.596	0.327	0.059	0.014	0.004	0.000	0.000
Cumulative	0.596	0.923	0.982	0.996	1.000	1.000	1.000

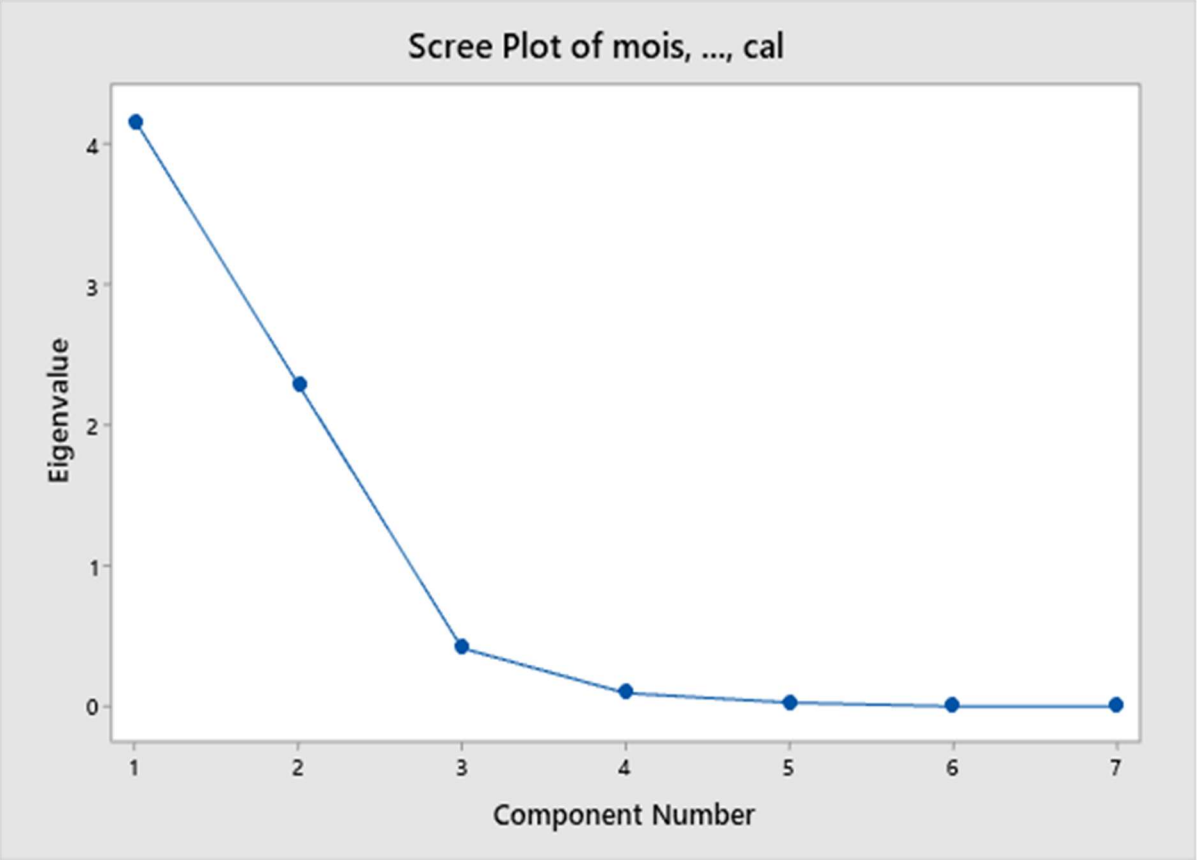
Here from eigen analysis we get an idea that by first Principal component 59% variation is explained , similarly by second and third principal component 32% and 5% variation is explained.

Eigenvectors

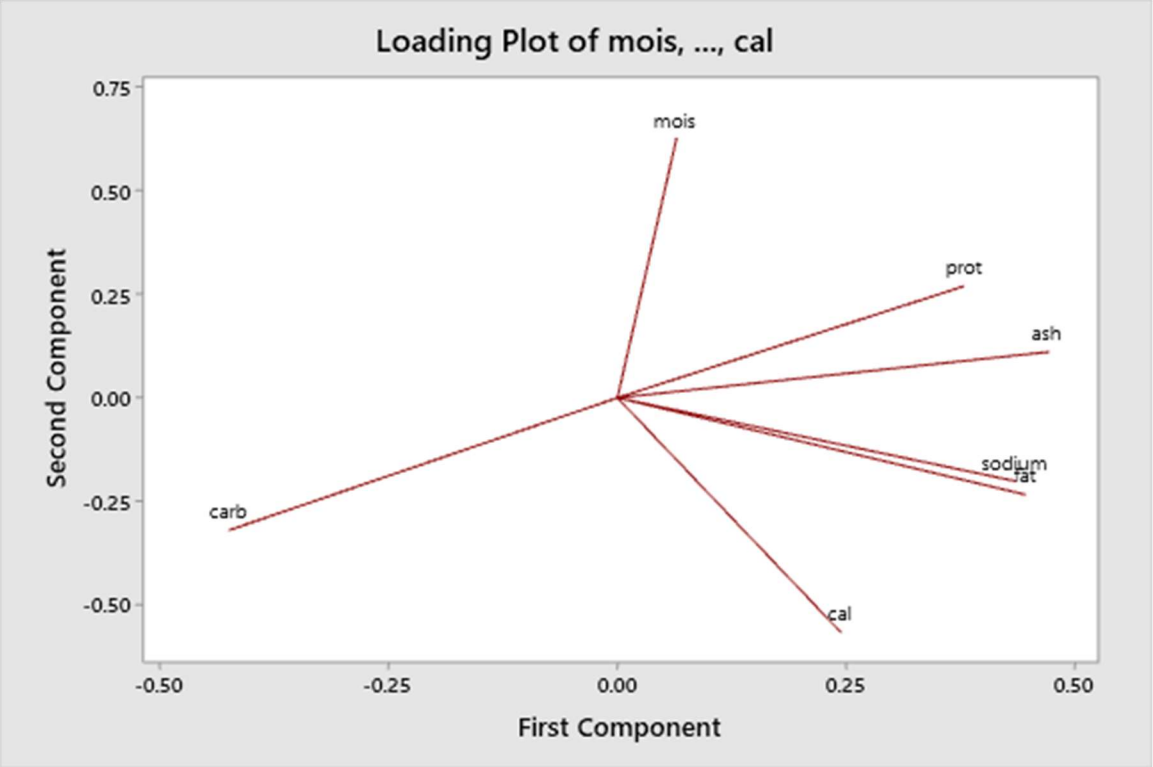
Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7
mois	0.065	0.628	-0.422	-0.221	-0.006	-0.446	-0.419
prot	0.379	0.270	0.746	-0.011	-0.388	0.000	-0.277
fat	0.447	-0.234	-0.199	-0.507	0.173	0.525	-0.378
ash	0.472	0.111	0.056	0.552	0.671	-0.059	-0.056
sodium	0.436	-0.202	-0.455	0.446	-0.603	-0.003	0.001
carb	-0.425	-0.320	0.052	0.334	0.007	0.001	-0.776
cal	0.244	-0.567	0.113	-0.279	0.078	-0.722	-0.012

The first principal component accounts for 59.6% of the total variation. The variables that correlate the most with this component .i.e first are fat ,ash , sodium. the first principal component correlated positively with this variables. And overall the 3 principal component explain the approx. 95% variation of total variabilitiy.

Scree plot:-



From scree plot we can see that bend occur at 3rd eigen value. So we can use first 3 princpal components.



Conclusion:-

Here we analysed nutrient content in pizza The data set of pizza contains measurements that capture the kind of things that make a pizza tasty which has8 variables but after analysing dataset through PCA technique we get 3 principal component which affect the most the taste of pizza.