This code utilizes principal component analysis (PCA) and k-means clustering to analyze a dataset containing information about compounds. The code reads in the dataset, handles missing values, transposes the data to make compounds the rows, and performs PCA to reduce the dimensionality of the data from the original number of variables to two principal components. K-means clustering is then applied to the PCA data to cluster the compounds into two groups but it can be changed as the client prefers. The clustering results are outputted to a CSV file with the compound names and a column indicating which cluster each compound belongs to.

A scatter plot of the PCA data is generated with points colored by cluster to provide a visual representation of the separation of the data into the two clusters. It is important to note that the choice of two or more clusters is arbitrary and may not be optimal for all datasets. The elbow method or silhouette score can be used to determine the optimal number of clusters.

This code can be used to gain insights into the relationships between compounds and identify patterns within the data. The clustering results can be further analyzed and used to make informed decisions about the compounds.

The input for the code is csv file name containing data like below form.

Window n	actuosus	yatensis	YPW6	AgN23	albidoflavus
1	-0.05474	0.018781	0.049352	0.025304	0.041463
2	-0.10123	0.086235	0.101894	0.075499	0.146006
3	-0.07755	0.106279	0.096836	0.141358	0.306251
4	-0.05115	0.076328	0.206957	0.354764	0.169545
5	-0.051	0.125694	0.165173	0.377739	0.13287
6	-0.09762	0.263822	0.259552	0.489818	0.120849
7	-0.00102	0.303554	0.191084	0.469997	0.171724
8	0.097837	0.437105	0.23806	0.408604	0.260929
9	0.24871	0.709126	0.366955	0.489751	0.336613
10	0.258954	0.652635	0.526539	0.605497	0.31978
11	0.29807	0.637229	0.575326	0.657353	0.354394
12	0.334291	0.609695	0.571792	0.615668	0.416169
13	0.225636	0.663564	0.579735	0.63597	0.452555
14	0.239119	0.629202	0.580693	0.683366	0.512505
15	0.228925	0.580133	0.586173	0.699887	0.555725
16	0.205608	0.680943	0.583382	0.673929	0.582638
17	0.241756	0.643818	0.575005	0.651515	0.638787
18	0.300384	0.700961	0.661489	0.6731	0.608297
19	0.338133	0.693263	0.665884	0.765116	0.585626
20	0.342357	0.703966	0.681749	0.771371	0.628021
21	0.349767	0.717114	0.712264	0.747387	0.718832
22	0.293533	0.682501	0.743592	0.794449	0.774917
23	0.308598	0.677318	0.740947	0.798961	0.865972

The output of the code are two csv files named “pca_output.csv” and “results.csv” and a visual graphical representation of compounds as dots on graph. “pca_output.csv” will contain the result of PCA and store PC1 in column 2 and PC2 in column 3, it will also give numbering to the compounds in column 4.

	A	B	C	D	E
1	Compounc	PC1	PC2	CompoundNumber	
2	actuosus	-9.17823	-2.12302	1	
3	yatensis	7.002976	-1.34947	2	
4	YPW6	-3.8771	0.298921	3	
5	AgN23	11.3625	0.47838	4	
6	albidoflavu	-5.31015	2.695195	5	

“results.csv” is same file as “pca_output.csv” but it will store cluster number in column 5 which is calculated by K-means clustering.

	A	B	C	D	E
1	Compounc	PC1	PC2	Compounc	Cluster
2	actuosus	-9.17823	-2.12302	1	0
3	yatensis	7.002976	-1.34947	2	1
4	YPW6	-3.8771	0.298921	3	0
5	AgN23	11.3625	0.47838	4	1
6	albidoflavu	-5.31015	2.695195	5	0
7					

In the graphical representation compounds with same colour represents same cluster and the numbering shown in graph is same as stored in csv files.