Wine Cultivar Clustering

Dominic Maranta

Abstract

- Wine Cultivar data set
- Sourced, cleaned, and modeled the data using Python.
- Unsupervised machine learning problem using clustering to identify cultivars
- Both k-means clustering and agglomerative clustering are highly effective when combined with PCA

Problem Statement

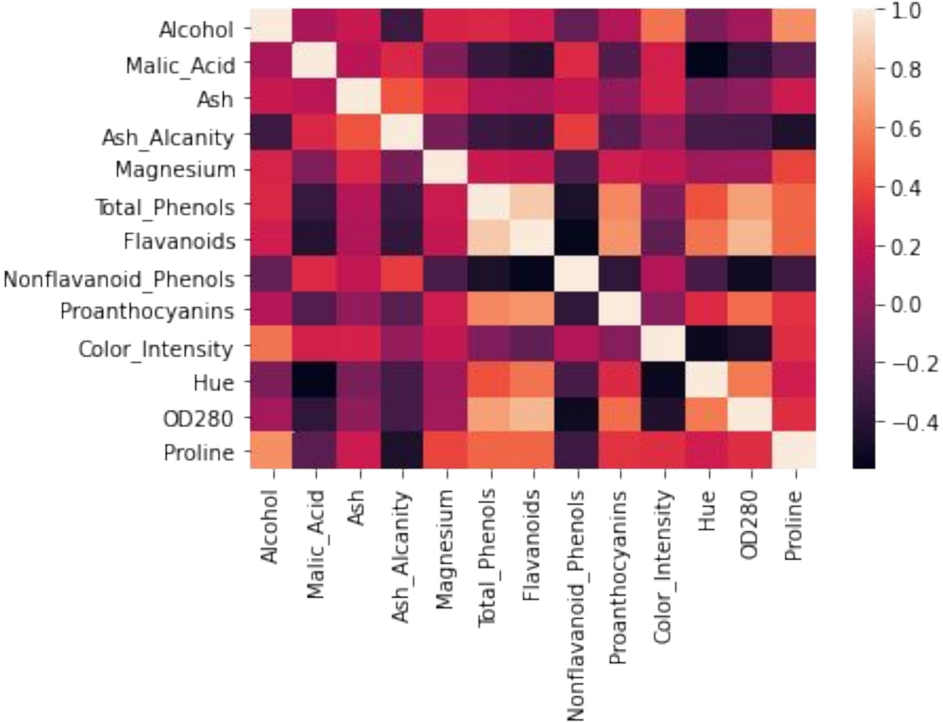
- Construct and train a model to cluster wine types
- How can we properly identify the cultivar a wine is from.
- Use EDA to explore and prepare the data.
- Compare and Evaluate machine learning clustering models.

Data Preprocessing

- Data: 13 variables with 178 instances
- 13 features identify characteristics including alcohol content, hue, and many more.
- No missing values in the 178 observations
- Clean variables and create numeric variables
- Perform EDA before modeling

Data Understanding

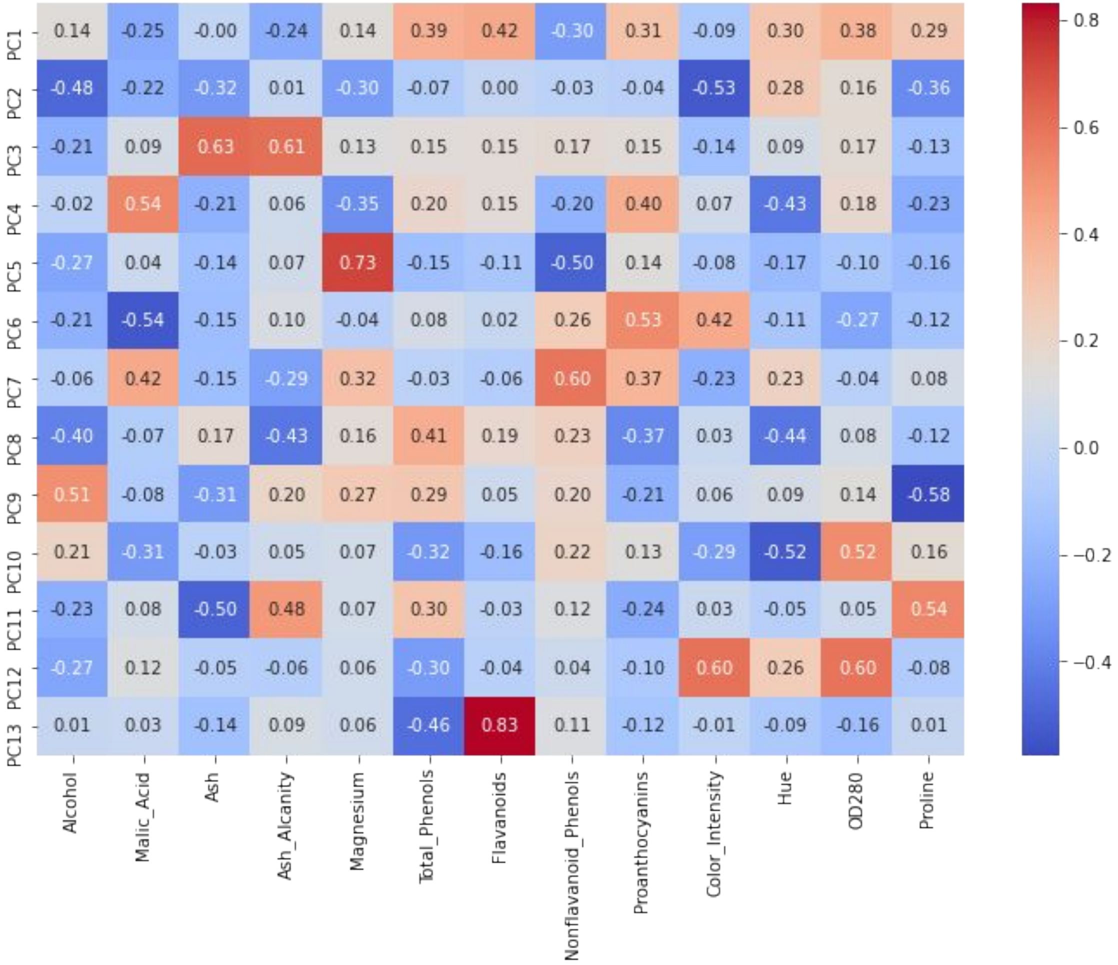
- Phenols, flavonoids, nonflavanoids, alcohol and ash alcanity, all produce strong correlation patterns, both inversely and directly
- Examining individual histograms from these 5 variables show distinct humps in the graphs, denoting potential differences in cultivars



Modeling

- Before clustering a key step is to standardize all the data.
- Apply a PCA transformation to reduce dimensions, multicollinearity, and assist with visualization.
- First 2 Principal components explain 47% of the variation.
- Phenols, flavonoids, nonflavanoids, and ash alcanity all important in the first PC.
- Alcohol and color intensity very important in the second PC.

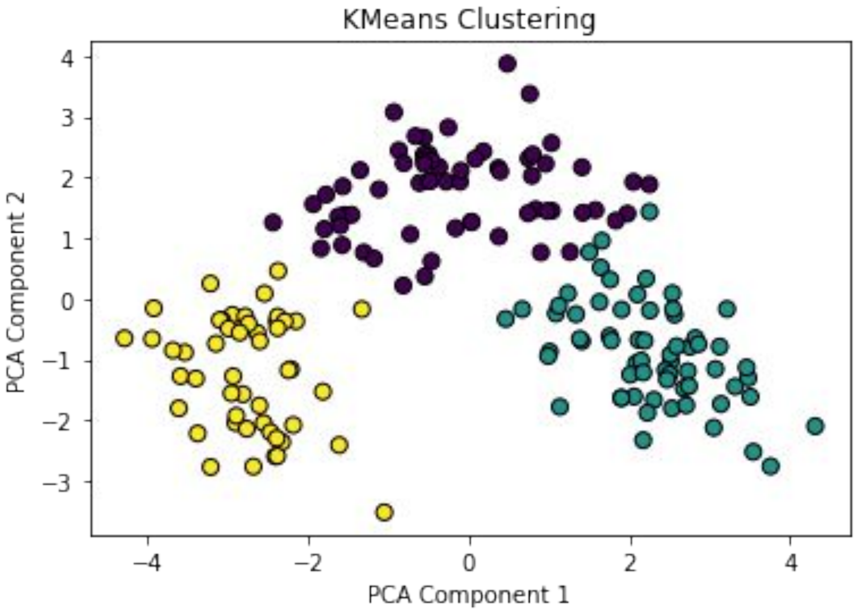
PCA:



	Alcohol	Malic_Acid	Ash	Ash_Alcanity	Magnesium	Total_Phenols	Flavanoids	Nonflavanoid_Phenols	Proanthocyanins	Color_Intensity	Hue	OD280	Proline
PC1	0.14	-0.25	-0.00	-0.24	0.14	0.39	0.42	-0.30	0.31	-0.09	0.30	0.38	0.29
PC2	-0.48	-0.22	-0.32	0.01	-0.30	-0.07	0.00	-0.03	-0.04	-0.53	0.28	0.16	-0.36
PC3	-0.21	0.09	0.63	0.61	0.13	0.15	0.15	0.17	0.15	-0.14	0.09	0.17	-0.13
PC4	-0.02	0.54	-0.21	0.06	-0.35	0.20	0.15	-0.20	0.40	0.07	-0.43	0.18	-0.23
PC5	-0.27	0.04	-0.14	0.07	0.73	-0.15	-0.11	-0.50	0.14	-0.08	-0.17	-0.10	-0.16
PC6	-0.21	-0.54	-0.15	0.10	-0.04	0.08	0.02	0.26	0.53	0.42	-0.11	-0.27	-0.12
PC7	-0.06	0.42	-0.15	-0.29	0.32	-0.03	-0.06	0.60	0.37	-0.23	0.23	-0.04	0.08
PC8	-0.40	-0.07	0.17	-0.43	0.16	0.41	0.19	0.23	-0.37	0.03	-0.44	0.08	-0.12
PC9	0.51	-0.08	-0.31	0.20	0.27	0.29	0.05	0.20	-0.21	0.06	0.09	0.14	-0.58
PC10	0.21	-0.31	-0.03	0.05	0.07	-0.32	-0.16	0.22	0.13	-0.29	-0.52	0.52	0.16
PC11	-0.23	0.08	-0.50	0.48	0.07	0.30	-0.03	0.12	-0.24	0.03	-0.05	0.05	0.54
PC12	-0.27	0.12	-0.05	-0.06	0.06	-0.30	-0.04	0.04	-0.10	0.60	0.26	0.60	-0.08
PC13	0.01	0.03	-0.14	0.09	0.06	-0.46	0.83	0.11	-0.12	-0.01	-0.09	-0.16	0.01

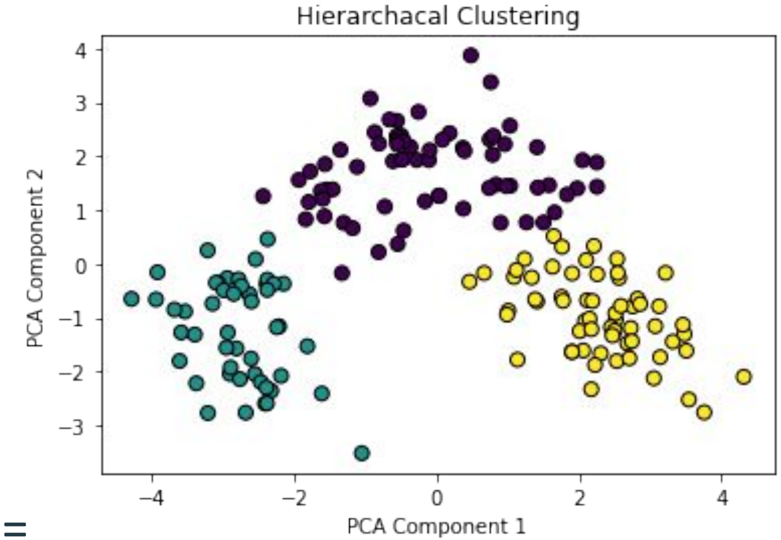
K Means Clustering

- K-means clustering performed optimally with 2 Principal components, compared to the raw data and an increased number of components
- Evaluated accuracy for three clusters = 96.6%



Agglomerative Clustering

- Agglomerative clustering performed optimally with 2 Principal components, compared to the raw data and an increased number of components
- Ward linkage methods performed much better than complete, single, and average
- Evaluated accuracy for three clusters = 96.6%



Outcomes

- Both agglomerative clustering and k-means clustering are effective, as long as PCA is used.
- New wine prediction is made easier with only wine features needed.
- Will allow for expanded understanding of how cultivars differ and which variables are important between cultivars.

Future Work

- Expand the dataset to include more features, observations, and cultivars.
- Apply supervised machine learning approaches to this data.
- Investigate the data with deep learning approaches.

Thank you!