Unsupervised Machine Learning Final Project: Wine Quality analysis

Main objective

The primary objective of this analysis is to explore the Wine Quality Dataset using unsupervised learning techniques, specifically clustering and dimensionality reduction.

- Identify distinct groups of wines based on their chemical properties.
- Reduce the dataset's dimensionality for better visualization and interpretation.
- Provide actionable insights for wine producers or quality control teams.

Unsupervised learning can help stakeholders understand patterns in wine characteristics, optimize production processes, and enhance product quality.

Dataset

❖ The dataset used is the **Wine Quality Dataset** which contains 12 attributes related to the physicochemical properties of wine samples and their quality ratings.

Key Features:
- **Fixed Acidity**: Concentration of tartaric acid.
- **Volatile Acidity**: Amount of acetic acid in wine.
- **Citric Acid**: Adds freshness to wine.
- **Residual Sugar**: Sugar left after fermentation.
- **Chlorides**: Salt content in wine.
- **Free Sulfur Dioxide & Total Sulfur Dioxide**: Preservatives affecting shelf life.
- **Density**: Density of the wine.
- **pH**: Acidity level of the wine.
- **Sulphates**: Contributes to flavor and preservation.
- **Alcohol**: Alcohol content in wine.
- **Quality (Target)**: Wine's quality score (3–8)

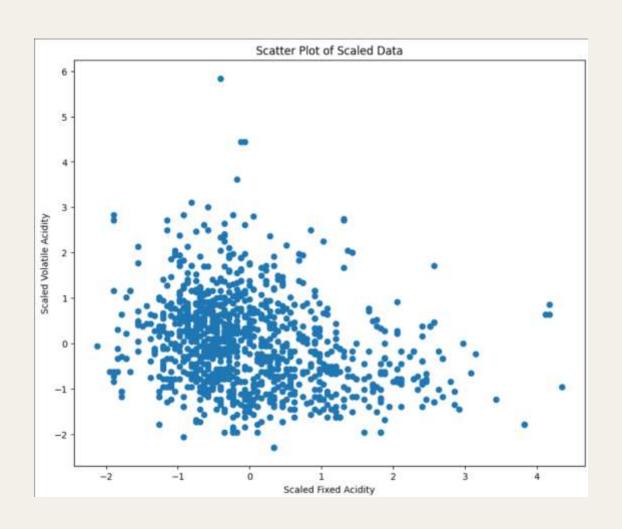|	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	Id
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	0
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5	1
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5	2
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6	3
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	4

Dataset

Summary:

- Number of Samples: 1,143
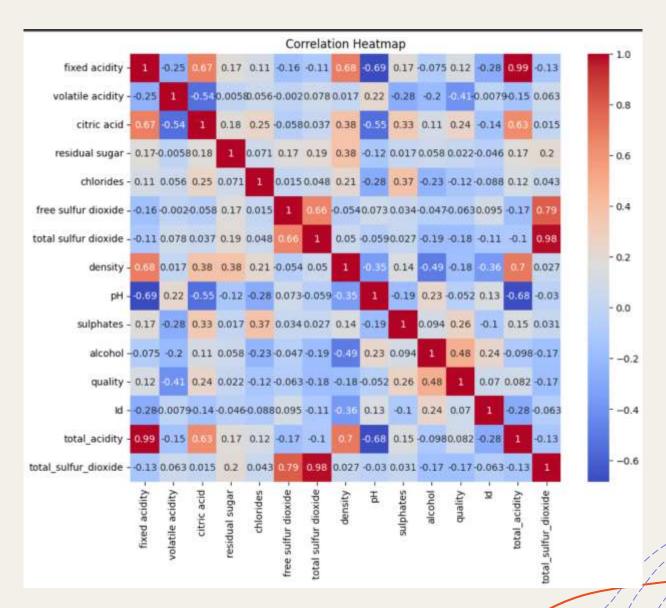- Number of Features: 12
- Target Variable: Quality

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	Id
count	1143.00	1143.00	1143.00	1143.00	1143.00	1143.00	1143.00	1143.00	1143.00	1143.00	1143.00	1143.00	1143.00
mean	8.31	0.53	0.27	2.53	0.08	15.62	45.91	1.00	3.31	0.66	10.44	5.66	804.97
std	1.75	0.18	0.20	1.36	0.06	10.25	32.78	0.00	0.16	0.17	1.08	0.81	464.00
min	4.60	0.12	0.00	0.90	0.01	1.00	6.00	0.99	2.74	0.33	8.40	3.00	0.00
25%	7.10	0.39	0.09	1.90	0.07	7.00	21.00	1.00	3.21	0.55	9.50	5.00	411.00
50%	7.90	0.52	0.25	2.20	0.08	13.00	37.00	1.00	3.31	0.62	10.20	6.00	794.00
75%	9.10	0.64	0.42	2.60	0.09	21.00	61.00	1.00	3.40	0.73	11.10	6.00	1209.50
max	15.90	1.58	1.00	15.50	0.61	68.00	289.00	1.00	4.01	2.00	14.80	8.00	1597.00

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	Id	total_acidity	total_sulfur_dioxide
0	7.40	0.70	0.00	1.90	0.08	11.00	34.00	1.00	3.51	0.56	9.40	5	0	8.10	45.00
1	7.80	0.88	0.00	2.60	0.10	25.00	67.00	1.00	3.20	0.68	9.80	5	1	8.68	92.00
2	7.80	0.76	0.04	2.30	0.09	15.00	54.00	1.00	3.26	0.65	9.80	5	2	8.56	69.00
3	11.20	0.28	0.56	1.90	0.07	17.00	60.00	1.00	3.16	0.58	9.80	6	3	11.48	77.00
4	7.40	0.70	0.00	1.90	0.08	11.00	34.00	1.00	3.51	0.56	9.40	5	4	8.10	45.00
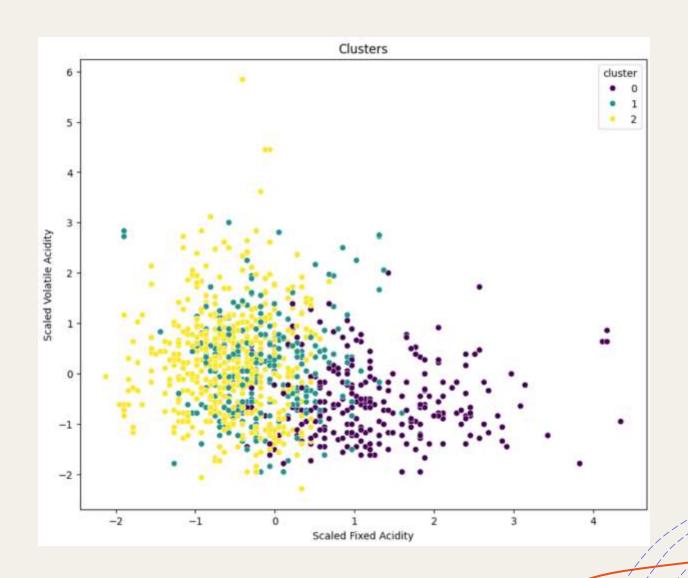
Data visualization

Data visualization

Correlation Heatmap

K-mean model

PCA

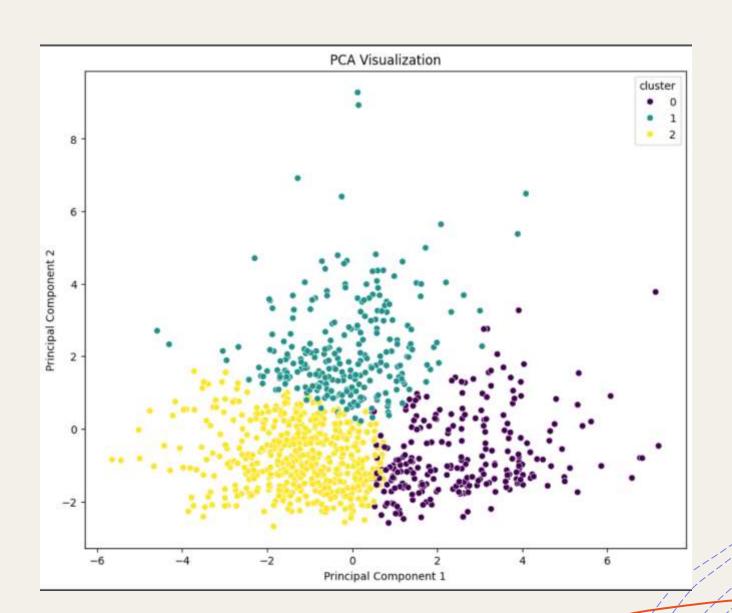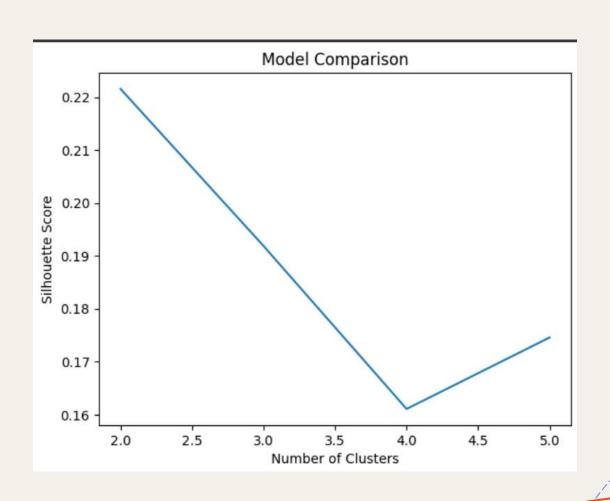PCA Visualization

Model comparison

Suggestions for Next Steps

1. **Feature Engineering**:
   Add new features like grape variety, region, or production year for deeper insights.
2. **Advanced Modeling**:
   Use t-SNE or UMAP for non-linear dimensionality reduction to capture complex patterns.
3. **Cluster Validation**:
   Validate clusters using external data like customer preferences or sales data.
4. **Business Applications**:
   Use cluster profiles to optimize wine production processes or target specific market segments.