CSCI-485

Assignment-2

Snehitha Gorantla

02-014-2025

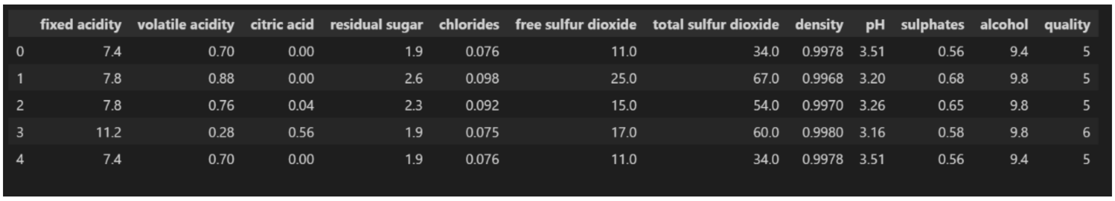
PCA AND DIMENSIONALITY REDUCTION

1. **Import Libraries:**

   Pandas for data manipulation, numpy for numerical operations, StandardScaler for data normalization, PCA for Principal Component Analysis, matplotlib.pyplot for plotting, mpl_toolkits.mplot3d for 3D plotting, seaborn for enhanced visualizations, and TSNE for t-SNE.

2. **Load and Prepare Data:**

   Load the red wine quality dataset from a URL using pandas. It then splits the data into features (everything except 'quality') and the target variable 'quality'.



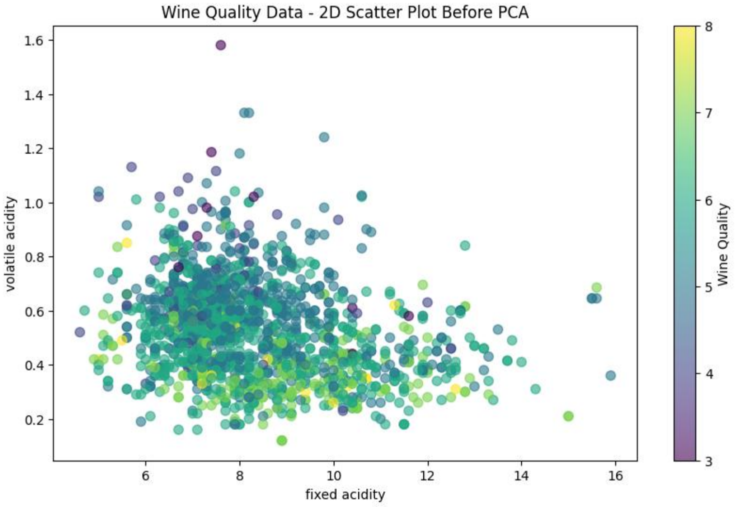
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970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

3. **Normalize Data:**

   The code uses StandardScaler to normalize the features. Normalization is important because PCA and t-SNE are sensitive to the scales of the features. It ensures that all features contribute equally to the analysis.

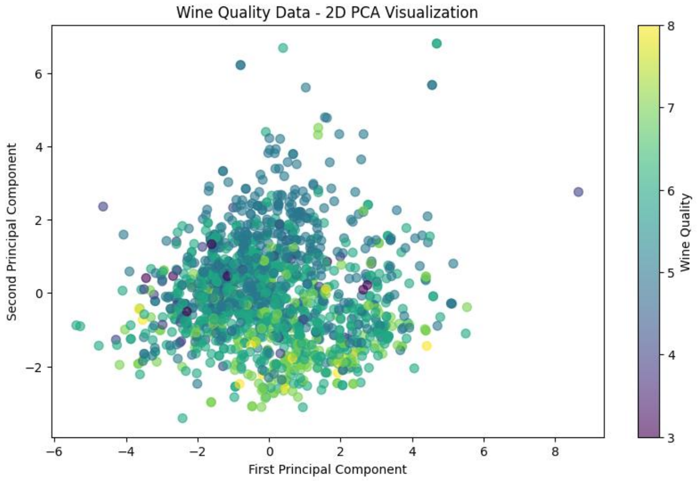
4. **2D Scatter Plot Before PCA:**

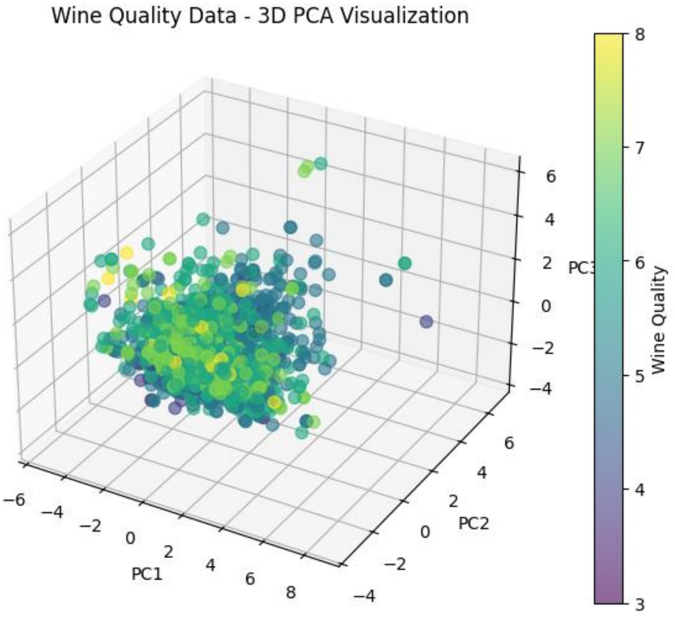
   Creates a 2D scatter plot of two of the original features, colored by wine quality. This gives a visual sense of the data distribution in the original feature space.



5. **PCA (2D and 3D):**

   - Applies PCA to the normalized features to reduce the dimensionality to 2 and 3 components.
   - Calculates and prints the explained variance ratio for the first two principal components. This tells you how much of the total variance in the data is captured by these components.
   - Creates 2D and 3D scatter plots of the PCA results, colored by wine quality. These plots visualize the data in the reduced-dimensional space

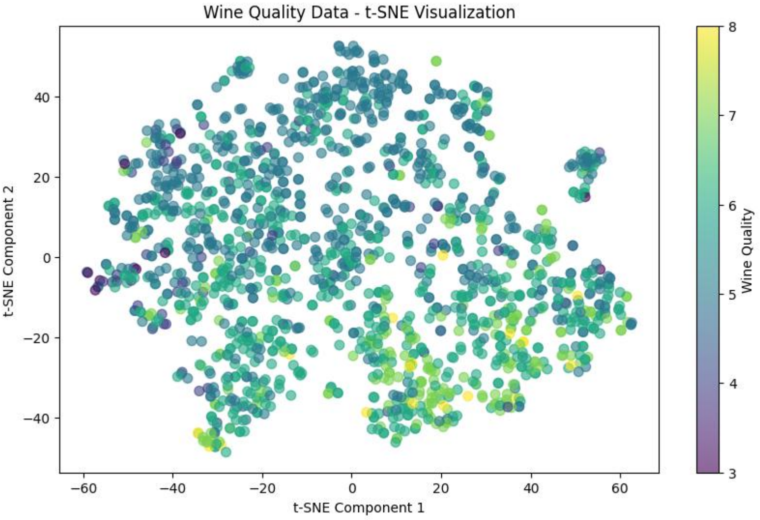


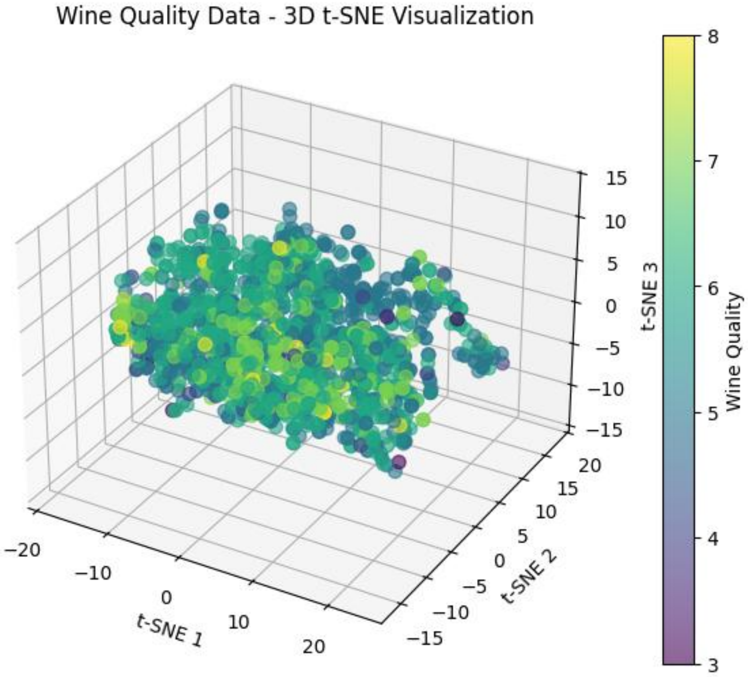


6. **t-SNE (2D and 3D):**

   - Applies t-SNE to the normalized features to reduce the dimensionality to 2 and 3 components. random_state is set for reproducibility.

- Creates 2D and 3D scatter plots of the t-SNE results, colored by wine quality. These plots visualize the data in the reduced-dimensional space, emphasizing local neighborhoods and clusters.





PCA vs. t-SNE for Wine Quality Data Visualization

This report compares the use of Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) for visualizing the wine quality dataset. Both methods aim to reduce the dimensionality of the data while preserving important information, but they approach this task differently and have distinct strengths and weaknesses.

PCA:

PCA is a linear dimensionality reduction technique that identifies the principal components of the data, which are orthogonal directions that capture the maximum variance. It's computationally efficient and provides insights into the variance explained by each component. However, PCA assumes linear relationships between features and may not capture complex non-linear structures in the data.

- **Pros:** Computationally fast, explains variance, identifies important linear combinations of features.
- **Cons:** Assumes linearity, may not be suitable for highly complex datasets.

t-SNE:

t-SNE is a non-linear dimensionality reduction technique that focuses on preserving the local structure of the data. It models the probability distribution of pairwise similarities between data points and aims to embed them in a lower-dimensional space while minimizing the difference between the high-dimensional and low-dimensional distributions. t-SNE is particularly effective at visualizing clusters and separating groups of data points. However, it's computationally intensive and the resulting visualizations can be sensitive to the choice of parameters (like perplexity). Also, the axes in t-SNE plots don't have a direct interpretation like principal components in PCA.

- **Pros:** Excellent for visualizing clusters and non-linear structures.
- **Cons:** Computationally intensive, sensitive to parameters, axes are not directly interpretable.

Key Observations from Visualizations:

- **PCA:** The 2D and 3D PCA plots show some separation between different quality levels, but the separation is not very distinct. The explained variance by the first two components is relatively low, suggesting that much of the variance is still lost in the 2D projection. The 3D plot provides a slightly better view, but still, the clusters overlap considerably.
- **t-SNE:** The t-SNE visualizations, both 2D and 3D, reveal more distinct clusters corresponding to different wine quality levels. The separation between clusters is noticeably better than in the PCA plots, indicating t-SNE's ability to capture the non-linear structure of the data and preserve local neighborhoods.

Trade-offs:

The choice between PCA and t-SNE depends on the specific goals of the analysis. If the primary goal is to understand the variance explained by different components and identify important linear combinations of features, PCA is a good choice. If the goal is to visualize clusters and separate groups of data points, especially in the presence of non-linear relationships, t-SNE is generally more effective, although it is computationally more expensive and requires careful parameter tuning. In this wine quality dataset, t-SNE proved to be more effective at revealing the underlying cluster structure related to wine quality.