ASSIGNMENT -2

Introduction:

It's an analysis of the UCI Wine Quality dataset using Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) for dimensionality reduction and visualization. The objective is to compare both methods in terms of interpretability, clustering performance, and how they handle high-dimensional data.

Dataset Overview

The dataset consists of two separate files:

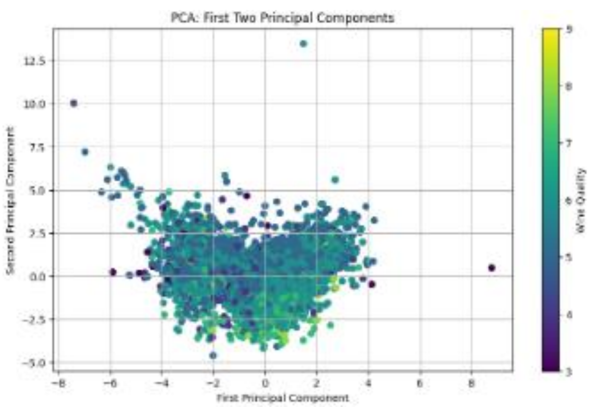
- winequality-red.csv (Red Wine Data)
- winequality-white.csv (White Wine Data)
- These datasets contain 11 physicochemical features along with a quality score (target variable). The datasets were combined for analysis.

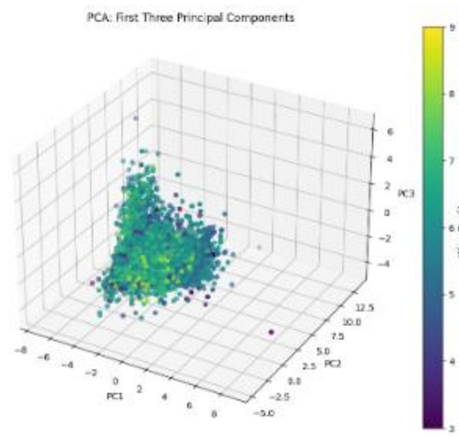
```
Red Wine Dataset (Normalized):
   fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
0      -0.528360          0.961877    -1.391472       -0.453218  -0.243707
1      -0.298547          1.967442    -1.391472        0.043416   0.223875
2      -0.298547          1.297065    -1.186070       -0.169427   0.096353
3       1.654856         -1.384443     1.484154       -0.453218  -0.264960
4      -0.528360          0.961877    -1.391472       -0.453218  -0.243707

   free sulfur dioxide  total sulfur dioxide   density        pH  sulphates  \
0            -0.466193             -0.379133  0.558274  1.288643  -0.579207
1             0.872638              0.624363  0.028261 -0.719933   0.128950
2            -0.083669              0.229047  0.134264 -0.331177  -0.048089
3             0.107592              0.411500  0.664277 -0.979104  -0.461180
4            -0.466193             -0.379133  0.558274  1.288643  -0.579207

    alcohol  quality
0 -0.960246        5
1 -0.584777        5
2 -0.584777        5
3 -0.584777        6
4 -0.960246        5

White Wine Dataset (Normalized):
   fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
0       0.172097         -0.081770     0.213280        2.821349  -0.035355
1      -0.657501          0.215896     0.048001       -0.944765   0.147747
2       1.475751          0.017452     0.543838        0.100282   0.193523
3       0.409125         -0.478657    -0.117278        0.415768   0.559727
4       0.409125         -0.478657    -0.117278        0.415768   0.559727

   free sulfur dioxide  total sulfur dioxide   density        pH  sulphates  \
0             0.569932              0.744565  2.331512 -1.246921  -0.349184
1            -1.253019             -0.149685 -0.009154  0.740029   0.001342
2            -0.312141             -0.973336  0.358665  0.475102  -0.436816
3             0.687541              1.121091  0.525855  0.011480  -0.787342
4             0.687541              1.121091  0.525855  0.011480  -0.787342

    alcohol  quality
0 -1.393152        6
1 -0.824276        6
2 -0.336667        6
```

Applying PCA

- PCA was applied to reduce the dataset to 2 and 3 principal components.
- Explained variance:
  - PC1: 27.54%
  - PC2: 22.67%
  - PC3: 14.15%
- The PCA scatter plot showed overlapping wine quality scores, indicating PCA does not effectively separate wine types.

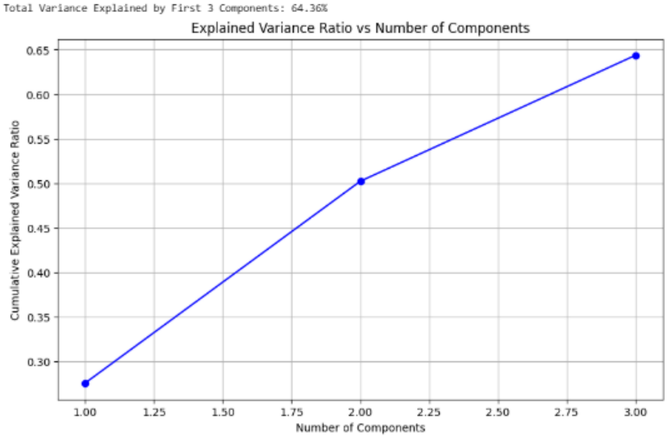
Transformed data using scatter plots (2D and 3D)



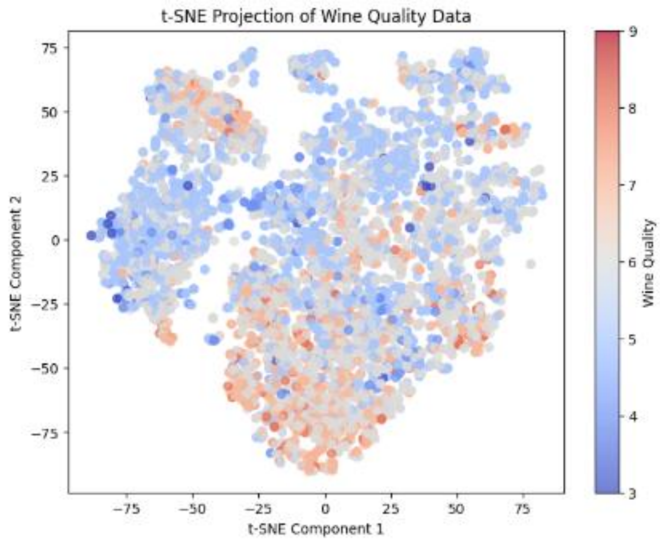


variance Explained by Each Principal Component

```
Variance Explained by Each Principal Component:
PC1: 0.2754 (27.54%)
PC2: 0.2267 (22.67%)
PC3: 0.1415 (14.15%)

Total Variance Explained by First 3 Components: 64.36%
```



Comparison with t-SNE



Conclusion

PCA is an effective tool for dimensionality reduction and feature selection, preserving global variance patterns while maintaining interpretability. However, it does not capture nonlinear relationships, making it less effective for clustering complex datasets.

t-SNE, in contrast, is better suited for clustering and visualizing hidden patterns by maintaining local relationships in the dataset. It is highly effective for grouping similar points together but is computationally expensive and lacks interpretability due to its nonlinear nature.

Ultimately, the choice between PCA and t-SNE depends on the task's goals. If interpretability and structured feature selection are important, PCA is preferred. If the focus is on discovering hidden clusters and nonlinear relationships, t-SNE is the better choice