Rucha Damle

(1.) Report based on principle component analysis. Principal Component Analysis (PCA) is a popular statistical method applied for dimensionality reduction and feature extraction. The goal of PCA is to explain most of the variability in a dataset with fewer variables than the original dataset.

Objective: In this report, I have applied the PCA to the wine quality dataset. The objective is to investigate how certain factors of chemical composition are related to each other and to the wine's quality rating.

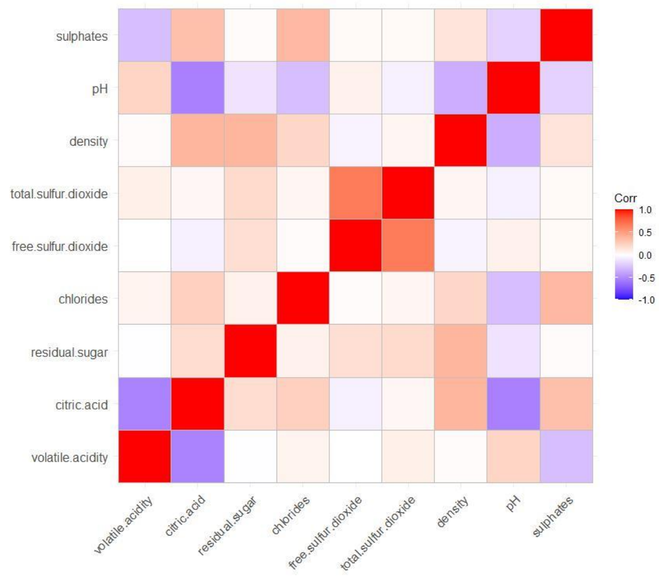
Methodology:

Exploring the data: The Wine Quality Dataset is a dataset of physical-chemical properties of red and white wine samples and their respective quality ordered from 0 to 10. Further, check the dataset for the null values.

Normalizing the data: The information in dataset has different scales and while performing PCA it can lead to a biased result. We ensure that each attributes has same level of contribution.

Correlation matrix of the data: The result of the correlation matrix can be interpreted as follow:

- The higher the value, the most positively correlated the two variables are.
- The closer the value to -1, the most negatively correlated they are.



Applying PCA: We can notice that 9 components have been generated. The cumulative proportion of first and second component is nearly 70% and first three components is 83%.

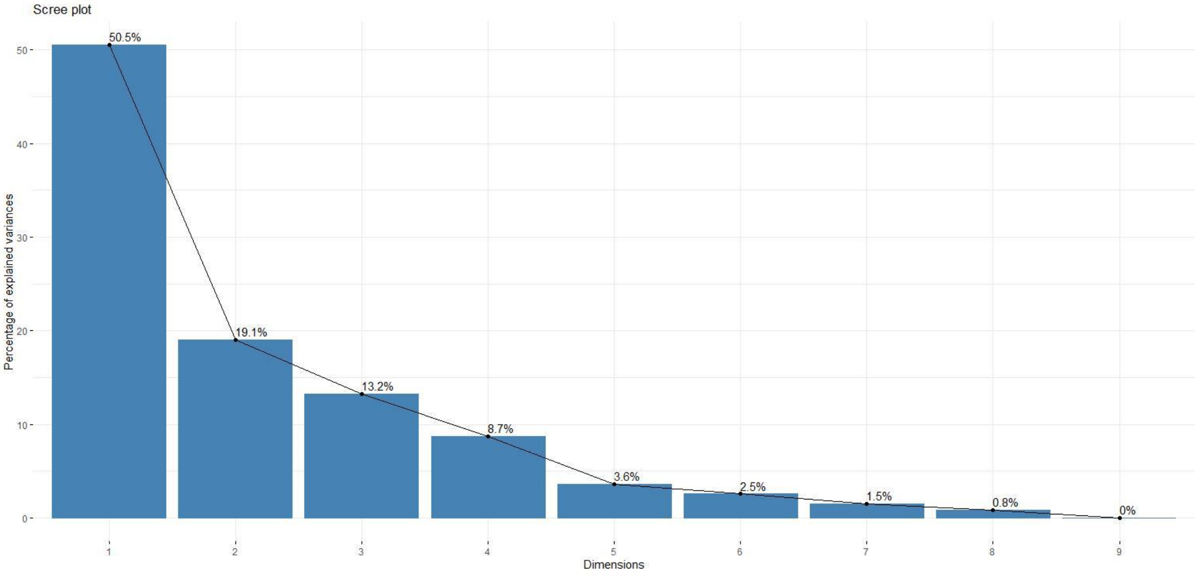
This implies that almost two-thirds of the data in the set of 9 variables can be represented by just two principal components.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	0.7891957	0.4846892	0.4036058	0.32805920	0.21118795
Proportion of Variance	0.5051674	0.1905428	0.1321237	0.08729116	0.03617463
Cumulative Proportion	0.5051674	0.6957101	0.8278338	0.91512498	0.95129961

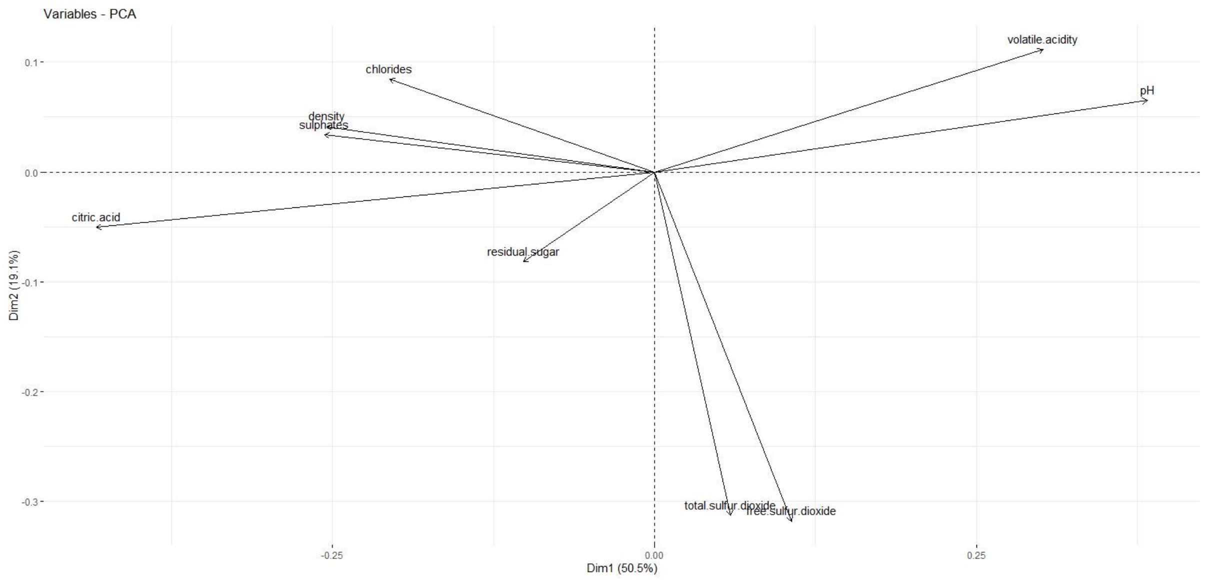
	Comp.6	Comp.7	Comp.8	Comp.9
Standard deviation	0.1772605	0.13565085	0.10109970	0
Proportion of Variance	0.0254853	0.01492488	0.00829021	0
Cumulative Proportion	0.9767849	0.99170979	1.00000000	1

Visualizing:

Our first approach is Scree plot. It is used to visualize the importance of each principal component and can be used to determine the number of principal components to retain. This plot shows the eigenvalues in a downward curve, from highest to lowest.



Second approach is biplot. It is possible to visualize the similarities and dissimilarities between the samples. From the following plot it is observed that all the variables are grouped together are positively correlated with each other. Then, the higher the distance between the variable and the origin, the better represented that variable is. Finally, variables that are negatively correlated are displayed to the opposite sides of the biplot’s origin.



Application:

Finding Patterns: We can find patterns in the chemical composition of wines by examining the primary components and the loadings that correspond to them. For instance, we might find that some mixes of chemical characteristics have a greater effect on wine's overall quality.

Reduction of Dimensionality: PCA assists in lowering the dataset's dimensionality while preserving the majority of its variation. This facilitates comprehension and interpretation by streamlining the data's analysis and visualization.

Quality Prediction: While PCA cannot predict anything directly, it can provide a reduced-dimensional representation of the data that can be utilized as input for machine learning models that can more accurately estimate the quality of wine.

Conclusion: PCA offers a strong tool for delving into and complex datasets such as the Wine Quality Dataset. Its implementation provides opportunity for additional research and analysis as well as insightful information about the factors impacting wine quality.

(2.) Report based on Factor analysis.

Factor analysis is a statistical technique used for dimensionality reduction and identifying the underlying structure (latent factors) in a dataset. It’s often applied in fields such as psychology, economics, and social sciences to understand the relationships between observed variables. Factor analysis assumes that observed variables can be explained by a smaller number of latent factors.

Objective: Investigating the underlying structure or dimensions inside the dataset's attributes and their connection to the wines' quality ratings is the aim of factor analysis applied to a wine quality dataset.

Methodology:

Exploring the data: A dataset containing information about all the chemical properties and the quality ratings is used for the analysis.

Unique variance: These values represent the unique variance in each observed variable that is not explained by the factors.

volatile.acidity	citric.acid	residual.sugar
0.005	0.275	0.872
chlorides	free.sulfur.dioxide	total.sulfur.dioxide
0.802	0.020	0.518
density	pH	sulphates
0.644	0.561	0.836

Loadings: These values represent the factor loadings for each observed variable on the extracted factors. Positive and high loadings indicate a strong relationship.

	Factor1	Factor2	Factor3
volatile.acidity	0.994		
citric.acid	-0.595		0.608
residual.sugar		0.190	0.301
chlorides			0.441
free.sulfur.dioxide		0.986	
total.sulfur.dioxide		0.681	0.127
density			0.596
pH	0.270		-0.604
sulphates	-0.301		0.268

SS Loadings: These are the sum of squared loadings for each factor, indicating the proportion of variance in the observed variables explained by each factor.
Proportion Var: This shows the proportion of total variance explained by each factor.
Cumulative Var: This shows the cumulative proportion of total variance explained as more factors are added.

	Factor1	Factor2	Factor3
SS loadings	1.510	1.481	1.474
Proportion Var	0.168	0.165	0.164
Cumulative Var	0.168	0.332	0.496

Test of the hypothesis: This section provides a chi-square test of whether the selected number of factors is sufficient to explain the variance in the data.

Test of the hypothesis that 3 factors are sufficient.

The chi square statistic is 263.05 on 12 degrees of freedom.

The p-value is 2.58e-49.

Application:

The primary applications are same as PCA which are understanding wine quality and dimension reduction.

Cluster analysis is one of the applications. The visualization of the transformed data enabled the identification of potential clusters or groups of wine samples with similar characteristics.

Conclusion:

The wine quality dataset's factor analysis uncovered underlying characteristics that affect perceived quality, including acidity, sweetness, and tannin levels. This study made it easier to reduce dimensionality and gave insightful information on how chemical attributes and quality ratings relate to one another. Overall, Factor Analysis serves as a powerful tool for exploring complex datasets and uncovering meaningful patterns in wine quality assessment.

(3.) Report based on canonical correlation analysis.

Canonical correlation analysis (CCA) is used to identify and measure the associations among two sets of variables. Canonical correlation is appropriate in the same situations where multiple regression would be, but where are there are multiple intercorrelated outcome variables.

Objective: The main objective of CCA is to find meaningful correlation between chemical properties of wines and quality ratings on the wine quality dataset.

Methodology:
Primary steps for analysis are same as previous techniques wiz. Exploring the dataset, preprocessing.
Applying CCA: CCA is applied to the pre-processed dataset using appropriate statistical software or libraries. This involves computing canonical variates and canonical correlations between the two sets of variables.
To determine the direction and intensity of the links between wine quality ratings and chemical attributes, canonical loadings and correlations are examined. The two sets of variables have a strong link, as indicated by significant canonical correlations.
Summary of CCA

```
        Length Class  Mode
cor      6     -none- numeric
xcoef   36     -none- numeric
ycoef   49     -none- numeric
xcenter  6     -none- numeric
ycenter  7     -none- numeric
```

CCA results for X

```
                              [,1]         [,2]          [,3]         [,4]         [,5]
fixed.acidity         0.0151129983 -0.002747175 -0.0024220413 -0.007071534  0.011754488
volatile.acidity      0.0039791213  0.044548960 -0.0734630741 -0.104203458  0.004131750
citric.acid           0.0020930050  0.068395672  0.0230216427  0.072300854 -0.039338579
residual.sugar        0.0055607180  0.005371927 -0.0095220968  0.002226067 -0.017719366
chlorides             0.0554710020 -0.092511713  0.4795052581 -0.342570077 -0.227954644
free.sulfur.dioxide  -0.0002366815  0.002529444  0.0005919518 -0.000337976  0.001311562
                             [,6]
fixed.acidity        -0.011907887
```

```
volatile.acidity     0.157520340
citric.acid          0.230105875
residual.sugar      -0.007033329
chlorides           -0.194556123
free.sulfur.dioxide -0.000567475
```

CCA results for Y

```
                              [,1]          [,2]          [,3]          [,4]          [,5]
total.sulfur.dioxide -8.666323e-05  9.094777e-04  4.160168e-05  5.367763e-05  1.445482e-04
density               1.236426e+01  4.900994e+00 -9.090534e+00 -1.294681e+00 -9.135633e+00
pH                   -1.026796e-01  1.342842e-02 -6.870228e-02 -1.411117e-02 -1.162437e-01
sulphates            -1.071846e-02 -5.580117e-03  1.475149e-01 -1.826556e-02 -1.092587e-01
alcohol               1.198400e-02  7.751857e-03 -1.328509e-02  1.192299e-02 -1.412388e-02
quality               3.928203e-04  9.806246e-04  1.970071e-03  2.611354e-02  1.374521e-02
Id                   -3.242203e-06  1.476356e-05  1.738818e-06  9.615776e-07 -1.850081e-05
                              [,6]          [,7]
total.sulfur.dioxide -1.854650e-05 -9.626301e-05
density              -5.761793e+00  1.533869e+00
pH                   -1.008529e-01 -6.016749e-02
sulphates             3.975081e-03 -2.172450e-02
alcohol               2.440422e-02 -2.358211e-03
quality              -3.258403e-02 -2.963884e-03
Id                   -1.473282e-05  6.308493e-05
```

Application:

Canonical Correlation Analysis can offer important insights into the variables influencing the perceived quality of wine in the context of the wine quality dataset. For instance, the investigation might show that specific pairings of chemical characteristics—for instance, increased sulphate and alcohol content combined with decreased acidity—are linked to better quality ratings. Wine makers may find this information useful in streamlining their production procedures in order to meet required quality requirements.

Conclusion:

Applying CCA to the wine quality dataset enables the identification of underlying relationships between wine quality ratings and chemical composition. Through the identification of noteworthy canonical correlations, this analysis offers valuable information that wine producers can apply to improve the quality of their wines.

In conclusion all of these techniques share similarities terms of dimensionality reduction and exploration of the data but differs in underlying assumptions and objectives.