Siddhi Gosavi

Multivariate statistical analysis

(1.) Report based on principle component analysis.

Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction and data visualization. It transforms a dataset with possibly correlated variables into a new set of uncorrelated variables called principal components. These components capture the directions of maximum variance in the data, allowing for simplification while retaining as much information as possible. PCA is widely used across various fields to uncover underlying patterns and relationships in high-dimensional datasets.

Objective:

The objective of this report is to perform Principal Component Analysis (PCA) on a wine quality dataset to explore its underlying structure and identify key factors contributing to wine quality.

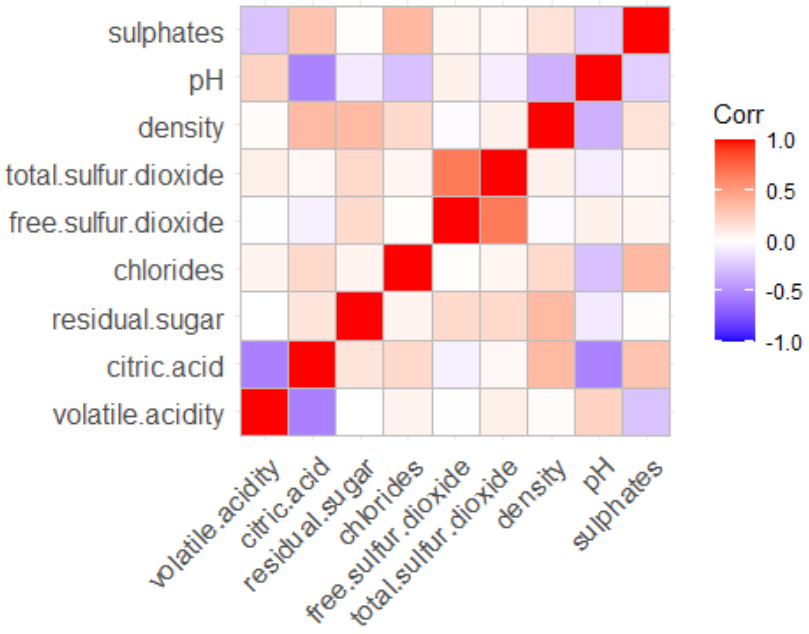
Methodology:

Data Collection: The wine quality dataset contains various physicochemical properties of wine, such as acidity, pH, alcohol content, and quality ratings provided by experts.

Data Preprocessing: The dataset is preprocessed by standardizing the variables to ensure that all features are on the same scale, which is a prerequisite for PCA.

Principal Component Analysis (PCA):

Compute the covariance matrix of the standardized data.



Interpretation: Analyze the loadings of the principal components to understand which variables contribute most to each component. Then we visualize the results to gain insights into the underlying structure of the wine quality dataset.

Application:

Applying PCA to the data in R and we are concluded with 9 components being generated

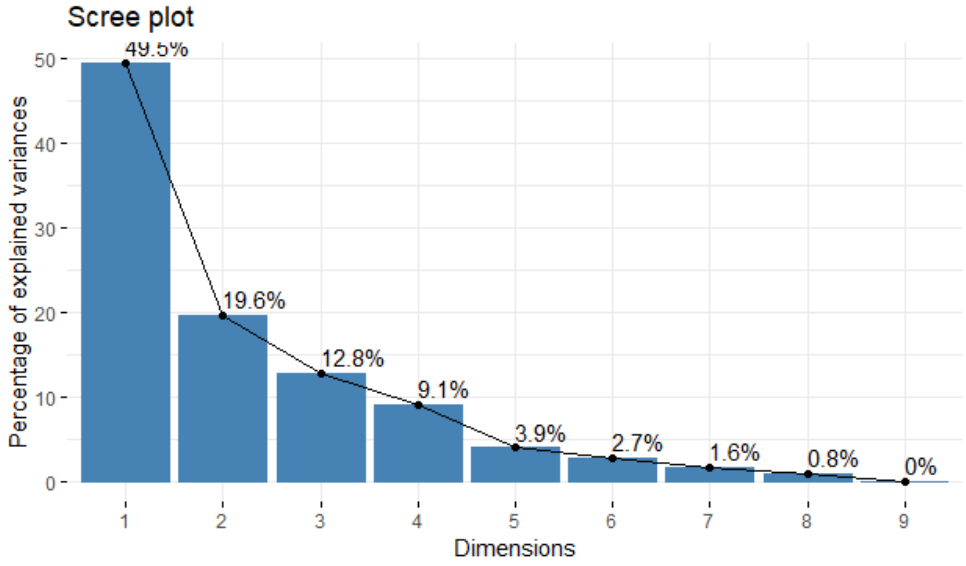
```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     C
omp.6
Standard deviation     0.7763802 0.4887568 0.3950594 0.33222983 0.21894505 0.182
30599
Proportion of Variance 0.4948076 0.1960980 0.1281186 0.09060763 0.03935118 0.027
28283
Cumulative Proportion  0.4948076 0.6909056 0.8190242 0.90963188 0.94898306 0.976
26589
                           Comp.7      Comp.8       Comp.9
Standard deviation     0.13891426 0.098057720 4.916603e-09
Proportion of Variance 0.01584095 0.007893164 1.984348e-17
Cumulative Proportion  0.99210684 1.000000000 1.000000e+00
> data.pca$loadings[, 1:2]
                          Comp.1      Comp.2
volatile.acidity       0.39651444  0.21477146
citric.acid           -0.55384419 -0.08627118
residual.sugar        -0.09831135 -0.20600508
chlorides             -0.24754177  0.21241897
free.sulfur.dioxide    0.11737166 -0.65258680
total.sulfur.dioxide   0.05848430 -0.63887654
density               -0.31157110  0.05383779
pH                     0.49320773  0.11938001
sulphates             -0.32796941  0.08767969
```

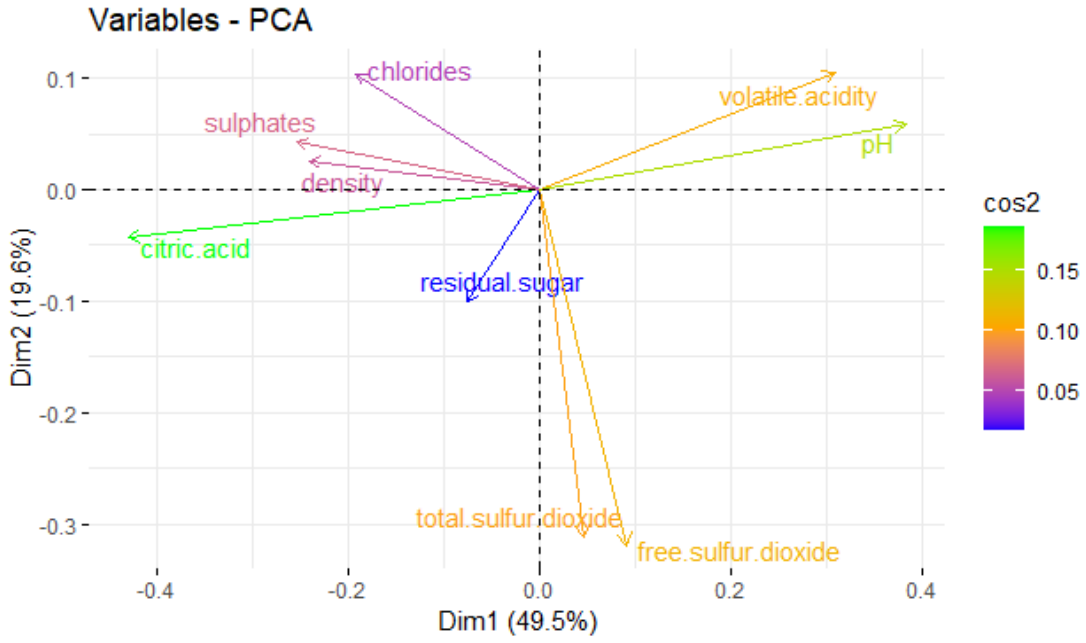
This implies that almost two-thirds of the data in the set of 9 variables can be represented by just two principal components- first and second. Application also tells us about the underlying patterns and is used for quality prediction.

Dimensionality Reduction: PCA will reduce the dimensionality of the wine quality dataset while preserving most of the information. By retaining a subset of principal components that capture the most variance in the data, we can represent the dataset in a lower-dimensional space.

Identification of Key Factors: PCA will help identify the key physicochemical properties that contribute to wine quality. By examining the loadings of the principal components, we can determine which variables have the greatest impact on the overall quality rating of the wines.

Visualization:





Conclusion:

Principal Component Analysis (PCA) applied to the wine quality dataset provides valuable insights into the underlying structure of the data and the factors influencing wine quality. By reducing the dimensionality of the dataset and identifying key variables, PCA enables us to better understand the relationships between physicochemical properties and wine quality ratings. These insights can inform winemaking processes and contribute to the production of higher-quality wines.

(1) Report on factor analysis

Factor Analysis (FA) is a statistical technique used to uncover underlying factors or latent variables that explain correlations among observed variables. It identifies common patterns in data by reducing the dimensionality and simplifying the structure, making it easier to interpret and understand complex datasets.

Objective:

This report aims to conduct Factor Analysis (FA) on a wine quality dataset to unveil underlying factors influencing wine quality and provide insights into the relationships between physicochemical properties and the expert-rated quality of wines.

Methodology:

Data Collection: The wine quality dataset comprises various attributes of wines, including acidity, pH levels, alcohol content, and quality ratings assigned by experts.

Data Preprocessing: Before conducting FA, the dataset is preprocessed. This involves handling missing values, ensuring data completeness, and standardizing the variables to have a mean of 0 and a standard deviation of 1, ensuring comparability across variables.

Factor Analysis (FA):

Factor Extraction: Determine the number of factors to extract, either based on theoretical considerations or statistical criteria such as the Kaiser criterion or scree plot. Then, apply factor extraction methods such as principal axis factoring or maximum likelihood estimation to identify underlying factors.
Factor Rotation: Rotate the extracted factors to achieve a simpler and more interpretable factor structure. Common rotation methods include Varimax, Promax, and Oblimin.
Interpretation: Interpret the rotated factor loadings to understand the relationship between the original variables and the extracted factors. Identify the variables with high loadings on each factor to label and interpret them accordingly.
Application:
Applying Factor Analysis to dataset in R gives us

Unique variance:

```
Uniquenesses:
    volatile.acidity          citric.acid       residual.sugar            chlori
des
               0.005                0.290                0.884                0.
827
 free.sulfur.dioxide total.sulfur.dioxide              density
pH
               0.039                0.496                0.646                0.
559
           sulphates
               0.843
```

Loadings:

```
Loadings:
                     Factor1 Factor2 Factor3
volatile.acidity      0.992
citric.acid          -0.611           0.580
residual.sugar                0.216   0.262
chlorides                             0.413
free.sulfur.dioxide           0.976
total.sulfur.dioxide          0.697   0.128
density                               0.593
pH                    0.293          -0.596
sulphates            -0.290           0.262
```

SS Loadings:

```
                Factor1 Factor2 Factor3
SS loadings       1.533   1.495   1.385
Proportion Var    0.170   0.166   0.154
Cumulative Var    0.170   0.336   0.490
```

Test of hypothesis:

```
Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 355.35 on 12 degrees of freedom.
The p-value is 1.04e-68
```

Identification of Latent Factors: FA uncovers latent factors that influence wine quality. By analyzing factor loadings, we can identify clusters of physicochemical properties that contribute to the overall quality of wines.

Dimension Reduction: FA reduces the dimensionality of the dataset by identifying a smaller number of factors that capture the variance in the original variables. This simplifies subsequent analyses and facilitates a deeper understanding of the dataset.

Conclusion:

Factor Analysis (FA) applied to the wine quality dataset reveals underlying factors that influence the quality ratings of wines. By extracting and interpreting these factors, we gain insights into the relationships between physicochemical properties and wine quality. FA

aids in dimensionality reduction and facilitates the identification of key factors contributing to wine quality, which can inform winemaking practices and contribute to the production of high-quality wines.

(2) Report based on canonical correlation analysis

Canonical Analysis is a multivariate statistical technique used to explore the relationship between two sets of variables. It seeks to find linear combinations of variables (canonical variates) that maximize the correlation between the two sets. Canonical analysis is particularly useful for identifying associations and dependencies between different sets of variables, facilitating a deeper understanding of their relationships.

Objective:

This report aims to conduct Canonical Analysis on a wine quality dataset to explore the relationship between physicochemical properties and expert-rated quality ratings of wines. The goal is to identify significant associations and dependencies between these two sets of variables.

Methodology:

Data Collection: The wine quality dataset contains various attributes of wines, including acidity, pH levels, alcohol content, and quality ratings provided by experts.

Data Preprocessing: Before conducting Canonical Analysis, the dataset is preprocessed to handle missing values and ensure data completeness. Additionally, both sets of variables (physicochemical properties and quality ratings) are standardized to have a mean of 0 and a standard deviation of 1 to ensure comparability.

Canonical Analysis:

Data Partitioning: The dataset is partitioned into two sets: predictor variables (physicochemical properties) and response variables (quality ratings).

Canonical Correlation Analysis: Canonical Correlation Analysis (CCA) is performed to identify linear combinations of predictor and response variables (canonical variates) that maximize the correlation between the two sets.

Interpretation: The canonical correlations and canonical loadings are examined to understand the strength and direction of the relationships between the two sets of variables. Significant canonical variates are identified, and their interpretations are provided based on the original variables.

Application:

We canonical correlation is applied to the dataset. 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

```
> summary(cca_result)
        Length Class  Mode
cor      6     -none- numeric
xcoef   36     -none- numeric
ycoef   36     -none- numeric
xcenter  6     -none- numeric
ycenter  6     -none- numeric
> # Canonical Correlation Coefficients
> cca_result$cor
[1] 0.9345108 0.7125917 0.5367743 0.4268398 0.1850280 0.1659524
```

CCA Results for X

```
> # Canonical Loadings for X
> cca_result$xcoef
                              [,1]          [,2]          [,3]          [,4]
fixed.acidity       -0.0131007129 -0.002065917 -0.0020145043 -0.0051154043
volatile.acidity    -0.0026595090  0.040969341 -0.0597674978 -0.0929292538
citric.acid         -0.0002141502  0.058160549  0.0170363227  0.0546435653
residual.sugar      -0.0046466691  0.003719081 -0.0080595742  0.0007428169
chlorides           -0.0526905389 -0.062448160  0.4145719684 -0.2789677208
free.sulfur.dioxide  0.0001958465  0.002131788  0.0005084632 -0.0001382334
                              [,5]          [,6]
fixed.acidity       -0.012905377  5.478320e-03
volatile.acidity     0.042560290 -1.238271e-01
citric.acid          0.112480443 -1.638335e-01
residual.sugar       0.010839773  1.099220e-02
chlorides            0.110895880  2.086209e-01
free.sulfur.dioxide -0.001134881 -8.510995e-07
quality              -1.993577e-02  2.146963e-02
```

CCA Results for Y

```
> # Canonical Loadings for Y
> cca_result$ycoef
                                [,1]          [,2]          [,3]          [,4]
total.sulfur.dioxide  8.200591e-05  0.0007719184  1.732424e-05  8.063418e-05
density              -1.085307e+01  3.3211597995 -8.146686e+00 -1.898735e+00
pH                    8.856989e-02  0.0148491977 -5.944438e-02 -2.238442e-02
sulphates             8.401829e-03 -0.0056775702  1.260793e-01 -1.876702e-02
alcohol              -1.002391e-02  0.0060564855 -1.148597e-02  9.608252e-03
quality              -3.342330e-04  0.0013589772  2.962052e-04  2.234066e-02
                              [,5]          [,6]
total.sulfur.dioxide -8.145362e-05 -4.110437e-05
density               4.707491e+00  6.407521e+00
pH                    6.209685e-02  1.237199e-01
sulphates             8.461502e-02  3.939217e-02
alcohol               2.071206e-02 -1.303448e-02
quality              -1.993577e-02  2.146963e-02
```

Identifying Associations: Canonical Analysis uncovers associations between physicochemical properties and wine quality ratings. By analyzing canonical correlations and loadings, we can identify which properties are most strongly related to high-quality wines.

Understanding Relationships: Canonical Analysis provides insights into the underlying relationships between predictor and response variables. This understanding can inform winemaking practices by highlighting the key factors contributing to wine quality.

Conclusion:

Canonical Analysis applied to the wine quality dataset reveals significant associations and dependencies between physicochemical properties and expert-rated quality ratings of wines. By identifying canonical variates that maximize the correlation between the two sets of variables, we gain insights into the underlying relationships and factors influencing wine quality. These insights can inform winemaking practices and contribute to the production of high-quality wines.

In conclusion, PCA, FA, and canonical analysis are powerful tools for analyzing multivariate data and extracting meaningful insights. By applying these techniques to the wine quality dataset, we can gain a deeper understanding of the factors influencing wine quality and potentially improve winemaking processes. These analyses demonstrate the versatility and utility of multivariate statistical techniques in various domains, including viticulture and oenology.