Multivariate Assignment

Report

Name: Sanskruti Sonawane

PRN No: 23060641041

Question 1:

We are performing principal component analysis (PCA) on the data “Quality of Wine”.

This data consists of 12 variables which are as follows:

```
library(readr)
> wine_data_csv <- read_csv("wine_data.csv.csv")
> View(wine_data_csv)
> names(wine_data_csv)
 [1] "fixed acidity"        "volatile acidity"     "citric acid"          "residual sugar"
 [5] "chlorides"            "free sulfur dioxide"  "total sulfur dioxide" "density"
 [9] "pH"                   "sulphates"            "alcohol"              "quality"
```

OBJECTIVE :
To find a new set of variables, smaller than the original set of variables such that maximum information of the data set is retained.

Applications of Principal Component Analysis (PCA) and its various purposes in data analysis and machine learning:

Dimensionality Reduction: PCA effectively simplifies complex datasets by transforming them into a smaller set of variables, known as principal components.
These components retain the essential information, enabling more manageable data analysis, particularly with high-dimensional datasets.

Data Visualization: By reducing the dimensionality of data, PCA facilitates visualization in lower dimensions. This aids in discerning patterns, clusters, and outliers that may not be readily apparent in the original high-dimensional space.

Feature Extraction: PCA generates new features from the original data, represented by the principal components.
These components capture the most significant variations in the data and are often independent, streamlining subsequent analysis, such as machine learning algorithms.

Data Preprocessing: PCA is a common preprocessing technique for diverse machine learning models . By reducing dimensionality and potentially eliminating redundant information, PCA enhances the efficiency and accuracy of machine learning algorithms.

Anomaly Detection: PCA assists in identifying outliers within datasets.
Since principal components encapsulate the most substantial variations, data points deviating significantly from these components are likely anomalies or outliers.

Methodology:

Executing Principal Component Analysis (PCA) in R Studio follows these steps:

Load Data: Commence by importing your dataset into R Studio using appropriate functions such as `read.csv()` for CSV files, ensuring data integrity.

Data Preprocessing: Prepare the data for PCA by addressing missing values, standardizing or scaling features if needed, and confirming that all variables are numeric.

Conduct PCA: Utilize the `prcomp()` function in R to conduct PCA on the dataset. This function computes the principal components along with their associated variance

Interpret Results: Review the summary of PCA results, which provides insights into the variance explained by each principal component and the cumulative proportion of explained variance. Additionally, visualize the results using scree plots to determine the optimal number of principal components.

Visualize Results: Employ biplots or other visualization methods to explore the relationships between variables and principal components, aiding in the identification of clusters or patterns within the data.

The output is as follows:

```
> my_pca
Standard deviations (1, .., p=12):
 [1] 1.7666827 1.4972916 1.2972739 1.1022799 0.9865412 0.8139977 0.7863319 0.7112472 0.6413326
[10] 0.5726425 0.4245216 0.2439629

Rotation (n x k) = (12 x 12):
                              PC1          PC2         PC3          PC4         PC5         PC6
fixed acidity         0.487883358  0.004173212  0.16482854  0.231098077 -0.07877938  0.05553130
volatile acidity     -0.265128984 -0.338967858  0.22708884 -0.041858245  0.29937933  0.29728700
citric acid           0.473335467  0.137358104 -0.10022856  0.056735802 -0.12014871  0.13663328
residual sugar        0.139154423 -0.167736336 -0.24362014  0.383037581  0.70936319  0.10931059
chlorides             0.197426792 -0.189788185  0.02660785 -0.654777820  0.26623723  0.33733656
free sulfur dioxide  -0.045880713 -0.259483136 -0.61611132  0.033711483 -0.15941286 -0.04264807
total sulfur dioxide  0.004066746 -0.363971374 -0.54073214  0.028459726 -0.21845284  0.11595360
density               0.370301191 -0.330780789  0.16872267  0.200693412  0.20879298 -0.42566742
pH                   -0.432720849  0.065440145 -0.06977056  0.005466181  0.25764682 -0.48035396
sulphates             0.254535354  0.109333620 -0.21291324 -0.560502367  0.21483493 -0.40374303
alcohol              -0.073176777  0.502708647 -0.22497138  0.091701428  0.25972635  0.39217625
quality               0.112488776  0.473166214 -0.22336929  0.036669226  0.13758414 -0.14183046
                             PC7         PC8          PC9         PC10         PC11         PC12
fixed acidity        -0.3072150 -0.20052866  0.17457815  0.182956014 -0.256437921 -0.638579761
volatile acidity     -0.6262337 -0.14612614  0.06022334 -0.155105626  0.377161229 -0.004661681
citric acid           0.2441486 -0.29633271  0.22097505 -0.346085556  0.624327833  0.070036908
residual sugar        0.2838543  0.17062614 -0.27818728  0.052236558  0.088077871 -0.183646374
chlorides             0.2305470  0.18692254  0.41993639  0.003862734 -0.208616670 -0.053931176
free sulfur dioxide  -0.1382604  0.01935607  0.31800012  0.585388580  0.237933171  0.051921666
total sulfur dioxide -0.1102087 -0.08989655 -0.12182276 -0.589188239 -0.355046842 -0.069792953
density              -0.1225465 -0.07950023  0.24907449 -0.043538098 -0.231453058  0.566644992
pH                    0.1856917 -0.31469303  0.46191598 -0.207609889 -0.005599072 -0.341230056
sulphates            -0.2334021 -0.27549158 -0.45268884  0.071918570  0.097637028 -0.067792979
alcohol              -0.1217188 -0.47118865  0.09652795  0.110605247 -0.319948696  0.317640325
quality              -0.4123879  0.61224719  0.24024309 -0.260239788  0.052465707 -0.008470476
> my_pca$rotation
                              PC1          PC2         PC3          PC4         PC5         PC6
fixed acidity         0.487883358  0.004173212  0.16482854  0.231098077 -0.07877938  0.05553130
volatile acidity     -0.265128984 -0.338967858  0.22708884 -0.041858245  0.29937933  0.29728700
citric acid           0.473335467  0.137358104 -0.10022856  0.056735802 -0.12014871  0.13663328
residual sugar        0.139154423 -0.167736336 -0.24362014  0.383037581  0.70936319  0.10931059
chlorides             0.197426792 -0.189788185  0.02660785 -0.654777820  0.26623723  0.33733656
free sulfur dioxide  -0.045880713 -0.259483136 -0.61611132  0.033711483 -0.15941286 -0.04264807
total sulfur dioxide  0.004066746 -0.363971374 -0.54073214  0.028459726 -0.21845284  0.11595360
density               0.370301191 -0.330780789  0.16872267  0.200693412  0.20879298 -0.42566742
pH                   -0.432720849  0.065440145 -0.06977056  0.005466181  0.25764682 -0.48035396
sulphates             0.254535354  0.109333620 -0.21291324 -0.560502367  0.21483493 -0.40374303
```

```
alcohol              -0.073176777  0.502708647 -0.22497138  0.091701428  0.25972635  0.39217625
quality               0.112488776  0.473166214 -0.22336929  0.036669226  0.13758414 -0.14183046
                            PC7          PC8          PC9         PC10         PC11         PC12
fixed acidity        -0.3072150 -0.20052866  0.17457815  0.182956014 -0.256437921 -0.638579761
volatile acidity     -0.6262337 -0.14612614  0.06022334 -0.155105626  0.377161229 -0.004661681
citric acid           0.2441486 -0.29633271  0.22097505 -0.346085556  0.624327833  0.070036908
residual sugar        0.2838543  0.17062614 -0.27818728  0.052236558  0.088077871 -0.183646374
chlorides             0.2305470  0.18692254  0.41993639  0.003862734 -0.208616670 -0.053931176
free sulfur dioxide  -0.1382604  0.01935607  0.31800012  0.585388580  0.237933171  0.051921666
total sulfur dioxide -0.1102087 -0.08989655 -0.12182276 -0.589188239 -0.355046842 -0.069792953
density              -0.1225465 -0.07950023  0.24907449 -0.043538098 -0.231453058  0.566644992
pH                    0.1856917 -0.31469303  0.46191598 -0.207609889 -0.005599072 -0.341230056
sulphates            -0.2334021 -0.27549158 -0.45268884  0.071918570  0.097637028 -0.067792979
alcohol              -0.1217188 -0.47118865  0.09652795  0.110605247 -0.319948696  0.317640325
quality              -0.4123879  0.61224719  0.24024309 -0.260239788  0.052465707 -0.008470476
>
```

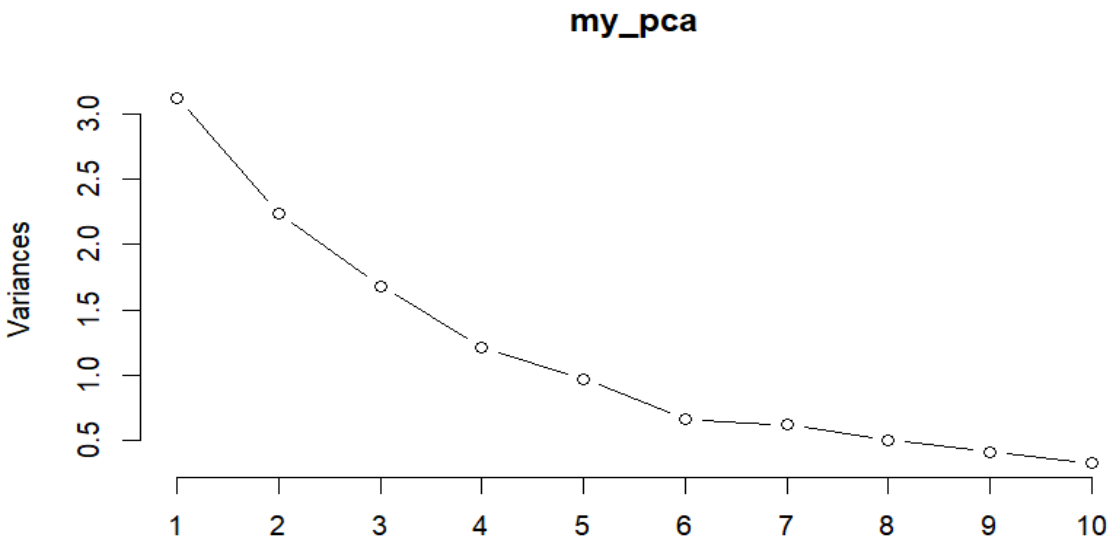
We can observe that:

Variables with larger absolute values in a given principal component contribute more to that component.

For example, in PC1, "fixed acidity," "citric acid," "density," and "pH" have relatively large absolute values, indicating they have a substantial influence on PC1.

Similarly, in PC2, "volatile acidity," "total sulfur dioxide," "density," "pH," and "alcohol" have larger absolute values, suggesting they contribute more to PC2.

The remaining PC's have high values only in one variable. Next, we plot these PC's in graph using the command "screeplot" in the following graph:



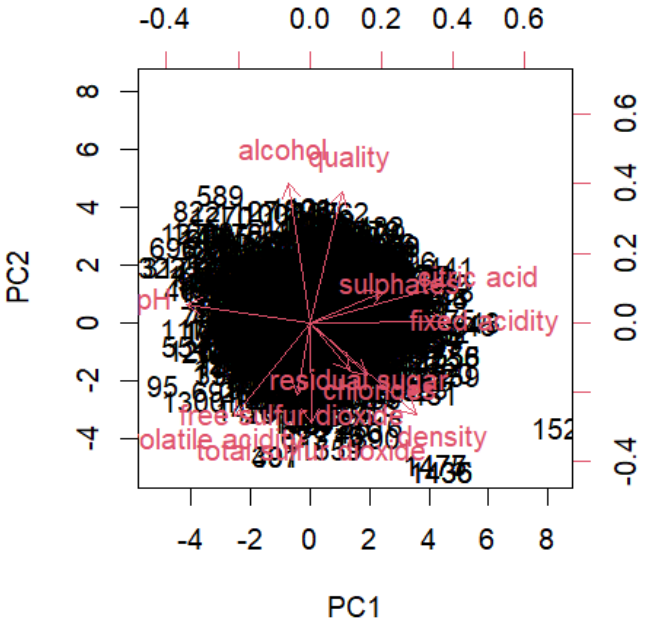
From the graph we observe that the starting 5 PC's have their eigen values greater than 1 we can use these 5 PC's for factor analysis.

Using the summary function, we can find the percentage of the total variation explained by these principal components:

1. PC1: 26.01%
2. PC2: 18.68%
3. PC3: 14.39%
4. PC4: 9.84%
5. PC5: 7.31%
6. PC6: 6.66%
7. PC7: 5.68%
8. PC8: 4.93%
9. PC9: 4.09%
10. PC10: 2.88%
11. PC11: 2.32%
12. PC12: 1.61%

These percentages represent the proportion of total variance explained by each principal component individually.

By using “Biplot” one can understand similar data patterns in regards to the variables in the original dataset



The graph indicates that some observations exhibit proximity to each other, suggesting similarity in data patterns concerning the variables in the original dataset. Additionally, it is evident that certain observations display stronger correlations with specific variables compared to others.

For example, we have many observations being very close to “fixed acidity” which signifies its high association with variable “fixed acidity”. Similarly, we also have observations being close to “sulphates” and “residual sugar” which also signifies more association with variable “hardness” and “residual sugar”

This graph also shows us how the variables are influenced by PC1 and PC2. If we consider the variable “density” we observe it strongly depends more on PC1 than on PC2 as its corresponding PC value is more.

Conclusion:

• We observe that some of the observations being close to each other which means that they have similar data patterns in regards to the variables in the original dataset and we also see that certain observations are more highly associated with certain variables than other observations.

• we observe that the starting 5 PC’s have their eigen values greater than 1 we can use these 5 PC’s for factor analysis.

• PC1 explains most of the variation in fixed acidity, citric acid, density,and pH.

• PC2 explains most of the variation in volatile acidity total sulfur dioxide, density, pH, and alcohol.

Question.2

OBJECTIVE:

To conduct Factor Analysis on the dataset, focusing only on Principal Components (PCs) with eigenvalues exceeding 1. The primary aim is to unveil latent factors that elucidate the relationships among multiple observed variables.

Applications of Factor Analysis in various fields:

Psychology and Personality Research: It serves as a fundamental tool in psychology, aiding researchers in validating psychological constructs such as intelligence, personality traits, or mental abilities. By identifying underlying factors, it explains scores on diverse psychological tests.

Market Research and Customer Segmentation: Marketers leverage factor analysis to comprehend customer preferences and segment their audience efficiently. Analyzing survey responses or product ratings helps identify underlying factors steering customer behavior, enabling grouping of customers with similar preferences.

Finance and Risk Management: In finance, factor analysis is instrumental in evaluating investment risk. By scrutinizing returns of multiple assets, it uncovers underlying factors influencing these returns, such as market risk or industry-specific factors.

Social Sciences and Education: Researchers in sociology, education, and other social sciences utilize factor analysis to explore relationships between various social indicators or educational assessments. It aids in identifying underlying factors like socioeconomic status or learning styles influencing these measures.

Data Mining and Machine Learning: Factor analysis contributes to data mining and machine learning as a dimensionality reduction technique. By decreasing the number of variables, it prepares data for further analysis and enhances the efficiency of machine learning algorithms.

By utilizing the Scree plot, we ascertain 5 PCs with eigenvalues exceeding 1. Consequently, we derive factors for these 5 PCs using the "factanal" function.

Methodology:

Read the Data: Read your dataset into RStudio using functions like read.csv() or read.table().

Data Preprocessing: Before conducting factor analysis, it's essential to preprocess the data, which may include handling missing values, scaling variables, or removing outliers.

Perform Factor Analysis: Use the factanal to conduct factor analysis.

Interpret Results: Once factor analysis is completed, you can examine the results, including factor loadings, communalities, eigenvalues, and scree plot.

Interpretation and Reporting: Finally, interpret the results of factor analysis and report your findings, including the extracted factors, their interpretations, and any implications for your research.

The output is as follows:

```
Call:
factanal(x = wine_data_csv, factors = 5, rotation = "varimax")

Uniquenesses:
       fixed acidity     volatile acidity          citric acid       resid
ual sugar
               0.005                0.503                0.232
0.680
           chlorides  free sulfur dioxide total sulfur dioxide
density
               0.864                0.541                0.005
0.005
                  pH            sulphates              alcohol
quality
               0.352                0.775                0.005
0.676

Loadings:
                     Factor1 Factor2 Factor3 Factor4 Factor5
fixed acidity         0.935  -0.160           0.162   0.264
volatile acidity     -0.177          -0.120  -0.663
citric acid           0.568                  0.645   0.149
residual sugar                0.157                  0.534
chlorides                            -0.262   0.245
free sulfur dioxide           0.669
total sulfur dioxide          0.984  -0.128           0.103
density               0.499          -0.465           0.724
pH                   -0.733           0.158  -0.256   0.134
sulphates                                    0.434   0.164
alcohol              -0.118           0.975   0.157
quality                      -0.127   0.429   0.347

                Factor1 Factor2 Factor3 Factor4 Factor5
SS loadings       2.044   1.500   1.484   1.352   0.977
Proportion Var    0.170   0.125   0.124   0.113   0.081
Cumulative Var    0.170   0.295   0.419   0.532   0.613

Test of the hypothesis that 5 factors are sufficient.
The chi square statistic is 948.81 on 16 degrees of freedom.
The p-value is 1.01e-191
```

Conclusion:

In the uniquenesses section, variables with higher uniqueness values indicate that a larger proportion of their variance is not accounted for by the extracted factors. In other words, variables with high uniqueness values are less well-explained by the factors and may have unique characteristics or measurement error. In your output:

Variables with high uniqueness values (greater than 0.5):

- Volatile acidity (0.503)

- Chlorides (0.864)

- Free sulfur dioxide (0.541)

- Sulphates (0.775)

- Quality (0.676)

Variables with low uniqueness values (closer to 0) are more well-explained by the factors.

Regarding the factor loadings, higher absolute values indicate stronger relationships between variables and factors. In your output, for each variable, the factor with the highest loading is considered the most influential factor for that variable. Looking at the loadings:

Factors with high loading values:

- Factor 1: Strong loadings for fixed acidity, citric acid, density, pH, and alcohol.

- Factor 2: Strong loading for free sulfur dioxide.

- Factor 3: Strong loadings for citric acid and total sulfur dioxide.

- Factor 4: Strong loadings for chlorides and density.

- Factor 5: Strong loading for alcohol.

Question.3.

Objective:

These interpretations are based on the absolute magnitude of the loadings. Higher absolute values indicate stronger associations between variables and factors.

Canonical correlation analysis (CCA) is a statistical technique used to assess the relationship between two sets of variables by identifying and quantifying their association beyond individual variable correlations.

Applications of CCA span various fields:

Neuroscience and Psychology: CCA helps analyze links between brain activity (measured via EEG or fMRI) and behavioral metrics, aiding in pinpointing brain regions associated with specific behaviors.

Chemoinformatics and Drug Discovery: CCA assists in understanding the correlation between the chemical structure of potential drugs and their biological activity, facilitating the development of new drugs.

Ecology and Environmental Science: CCA helps in deciphering the relationship between environmental factors (e.g., pollution levels) and ecological responses (e.g., species abundance), thus identifying significant environmental influences on ecological patterns.

Finance and Risk Management : CCA aids in analyzing the connection between risk factors (e.g., market volatility) and asset returns, assisting investors in comprehending how different risk factors affect portfolio performance.

Marketing and Customer Relationship Management (CRM): CCA assists in analyzing relationships between customer demographics, purchase history, and marketing campaign effectiveness, enabling companies to identify customer segments responsive to specific marketing messages and refine marketing strategies.

Methodology:

1. Data is divided into two sets: X and Y.

2. The "cancor" function is employed to calculate the canonical correlation between these two sets.

The output is as follows:

```
> # Canonical Correlation Coefficients
> cca_result$cor
[1] 0.9345108 0.7125917 0.5367743 0.4268398 0.1850280 0.1659524
> cca_result$xcoef
                              [,1]         [,2]          [,3]          [,4]
[,5]
fixed acidity       -0.0131007129 -0.002065917 -0.0020145043 -0.0051154043
-0.012905377
volatile acidity    -0.0026595090  0.040969341 -0.0597674978 -0.0929292538
0.042560290
citric acid         -0.0002141502  0.058160549  0.0170363227  0.0546435653
0.112480443
residual sugar      -0.0046466691  0.003719081 -0.0080595742  0.0007428169
0.010839773
chlorides           -0.0526905389 -0.062448160  0.4145719684 -0.2789677208
0.110895880
free sulfur dioxide  0.0001958465  0.002131788  0.0005084632 -0.0001382334
-0.001134881
                              [,6]
fixed acidity         5.478320e-03
volatile acidity     -1.238271e-01
citric acid          -1.638335e-01
residual sugar        1.099220e-02
chlorides             2.086209e-01
free sulfur dioxide  -8.510995e-07
> cca_result$ycoef
                                [,1]          [,2]          [,3]          [,
4]          [,5]
total sulfur dioxide  8.200591e-05  0.0007719184  1.732424e-05  8.063418e-
05 -8.145362e-05
density              -1.085307e+01  3.3211597995 -8.146686e+00 -1.898735e+
00  4.707491e+00
pH                    8.856989e-02  0.0148491977 -5.944438e-02 -2.238442e-
02  6.209685e-02
sulphates             8.401829e-03 -0.0056775702  1.260793e-01 -1.876702e-
02  8.461502e-02
```

```
alcohol              -1.002391e-02  0.0060564855 -1.148597e-02  9.608252e-
03  2.071206e-02
quality              -3.342330e-04  0.0013589772  2.962052e-04  2.234066e-
02 -1.993577e-02
                                [,6]
total sulfur dioxide -4.110437e-05
density               6.407521e+00
pH                    1.237199e-01
sulphates             3.939217e-02
alcohol              -1.303448e-02
quality               2.146963e-02
```

In the set of canonical correlation coefficients obtained from the analysis, each value signifies the strength of association between different pairs of linear combinations of variables from sets X and Y. Specifically, the first coefficient denotes the correlation between the first linear combination of variables in set X and the first linear combination in set Y, and this pattern continues for the subsequent coefficients up to the fifth.

Observing these coefficients, it's notable that the first canonical coefficient stands out with the highest value (0.93451) compared to the others. This indicates a robust relationship between the first linear combinations of variables in set X and Y, suggesting a significant correlation between these two sets of variables.
If we take the X coefficient and Y coefficient part, we observe that there is a strong relationship between first set of variables ("fixed acidity","volatile acidity", "citric acid", "residual sugar", "free sulfur dioxide", “chlorides”) and the second set of variables ("total sulfur dioxide","density", "ph", "sulphates", "alcohol",”quality”). This suggests that wine with higher fixed acidity, more volatile acidity, high contents of citric acid or residual sugar or free sulfur dioxide tend to have lower total sulfur dioxide, low contents of density or ph, less sulphates , less alcohol and less quality.

This interpretation is supported by the canonical loadings, which show that the variables in the first set are negatively correlated with the variables in the second set.

Conclusion:

• Strong relationship between the first linear combinations in “X” and the first linear combination in “Y”.

• Water with higher fixed acidity, more volatile acidity , high contents of citric acid, residual sugar, free sulfur dioxide, chlorides tend to have lower total sulfur dioxide, less density, low ph, less sulphates, less alcohol and quality.