Factor Analysis on Quality of Red Wine

S/18/837

April 7, 2024

1 Introduction

Nowadays wine is increasingly enjoyed by a wider range of consumers. To support its growth, the wine industry is investing in new technologies for both wine making and selling processes. Wine certification and quality assessment are key elements within this context. Certification prevents the illegal adulteration of wines and assures quality for the wine market.

The quality assessment of red wine is influenced by a Many different physical and chemical characteristics that contribute to its sensory perception. Understanding the intricate relationship between these attributes and the perceived quality of wine is paramount for winemakers and viticulturists aiming to enhance product excellence and consumer satisfaction.

In this exploratory data analysis, We explore factor analysis to identify what factors determine red wine quality. Leveraging a dataset compiled by Cortez and Reis. in 2009, we scrutinize eleven physical and chemical variables alongside one output variable, quality, rated on a scale from 0 to 10 based on sensory evaluations. Through factor analysis, we aim to find factors driving wine quality and illuminate insights vital for refining production processes and optimizing sensory experiences. This study's overarching purpose is to shed light on the intricate interplay between physical and chemical attributes and sensory perceptions, ultimately enhancing our understanding of the nuanced nature of red wine quality.

2 Methodology

2.1 Dataset Description:

In this Data Analysis on **Wine Quality** data set which has two dataset, related to red and white vinho verde wine samples, from the north of Portugal. The goal is to model wine quality based on physicochemical tests. The dataset utilized in this analysis comprises eleven physicochemical attributes alongside one output variable, quality, representing sensory evaluations of red wine. The physicochemical attributes include fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol content. Each attribute provides insight into different aspects of the chemical composition and properties of red wine.

Variable	Type	Description
fixed acidity	Continuous	Concentration of non-volatile acids in wine
volatile acidity	Continuous	Concentration of volatile acids in wine
citric acid	Continuous	Concentration of citric acid present in wine
residual sugar	Continuous	Amount of sugar remaining in the wine after fermentation
chlorides	Continuous	Concentration of chlorides in wine
free sulfur dioxide	Continuous	Concentration of free sulfur dioxide in wine
total sulfur dioxide	Continuous	Total amount of sulfur dioxide
density	Continuous	Density of the wine
pH	Continuous	Acidity or Alkalinity of the wine
sulphates	Continuous	Concentration of sulfates in wine
alcohol	Integer	Percentage of alcohol by volume in the wine
quality	Categorical	Overall quality of the wine

Table 1: **Variables Table**

2.2 Statistical Methods Employed:

Principal Component Analysis (PCA), Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) were employed as the statistical methods in this analysis.

Principle Component Analysis (PCA) is a dimensionality reduction technique used to identify patterns and relationships in high-dimensional data by transforming variables into a smaller set of uncorrelated variables called principal components. In the context of this analysis, PCA was utilized to reduce the dimensionality of the dataset while preserving as much of the original variability as possible. This facilitated the identification of the most influential physicochemical attributes contributing to red wine quality.

Factor Analysis (FA) is a statistical method used to uncover underlying

latent factors or dimensions that explain the correlations among observed variables. In this analysis, FA was employed to explore the complex relationships between the physicochemical attributes and the quality of red wine. By extracting factors that capture the shared variance among the attributes, FA facilitated the identification of the key factors driving variations in red wine quality scores.

Confirmatory Factor Analysis (CFA) is a statistical technique used to validate the underlying structure of a set of observed variables by confirming whether they reflect the hypothesized latent constructs (factors). Unlike Exploratory Factor Analysis (EFA), where the structure of the factors is derived from the data, CFA tests a pre-specified model of how the observed variables are related to the latent factors.

By using PCA, EFA and CFA, this methodology aimed to provide a comprehensive understanding of the interplay between physicochemical attributes and red wine quality, offering insights valuable for quality assessment and improvement in wine making process.

3 Results and Discussion

3.1 Exploratory Factor Analysis (EFA) :

The results obtained from the principal component analysis (PCA) and factor analysis (FA) provide valuable insights into the relationship between physicochemical attributes and the quality of red wine.

By employing PCA for the Wine quality data set, First we can identify the numbers of principle components that the majority of the variance in the dataset could be explained.

	eigenvalue <dbl>	variance.percent <dbl>	cumulative.variance.percent <dbl>
Dim.1	3.09913244	28.1739313	28.17393
Dim.2	1.92590969	17.5082699	45.68220
Dim.3	1.55054349	14.0958499	59.77805
Dim.4	1.21323253	11.0293866	70.80744
Dim.5	0.95929207	8.7208370	79.52827
Dim.6	0.65960826	5.9964388	85.52471
Dim.7	0.58379122	5.3071929	90.83191
Dim.8	0.42295670	3.8450609	94.67697
Dim.9	0.34464212	3.1331102	97.81008
Dim.10	0.18133317	1.6484833	99.45856

Figure 1: Eigenvalues

Since there are 5 eigenvalues closer to 1 and their cumulative total variance is approximately 80% , first 5 PC's can be retained

From the obtained results,

- The first Principle component explains 28.18% of the total variance.

- The second Principle component explains 17.51% of the total variance.
- The third Principle component explains 14.10% of the total variance.
- The fourth Principle component explains 11.03% of the total variance.
- The fifth Principle component explains 8.72% of the total variance.
- Therefore the total proportion of the variance explains by the first 4 principle components is 79.52%, which is approximately 80%.

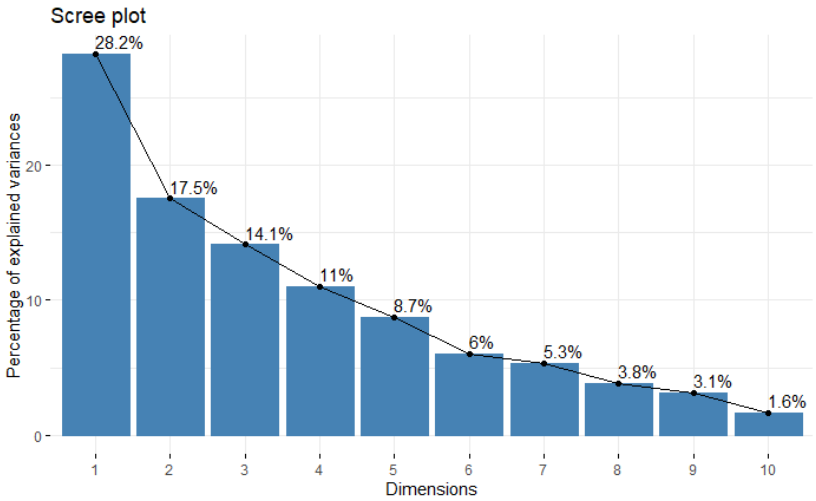

Figure 2: Scree plot

According to the scree plot, no steep drop is visible after the fifth PC. Therefore, we can retain the first 5 PCs after PCA.

Factor analysis

Factor Analysis elucidated the underlying latent factors that influence the observed correlations among the physicochemical attributes and quality scores. By extracting these factors, we were able to identify distinct dimensions that capture the shared variance among the variables.

To assess the adequacy of our factor analysis model, we conducted a hypothesis test comparing the fit of a model with five factors against the alternative hypothesis that more factors are needed.

Hypothesis Testing

H_0 : 5 factors are sufficient vs H_1 : More factors needed

```
factanal(x = wine_data[, c(1:11)], factors = 5, rotation = "varimax",     method = "principal")

Uniquenesses:
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar            chlorides
               0.005                0.567                0.146                0.680                0.857
 free.sulfur.dioxide total.sulfur.dioxide              density                   pH            sulphates
               0.540                0.005                0.005                0.356                0.824
             alcohol
               0.005

Loadings:
                     Factor1 Factor2 Factor3 Factor4 Factor5
fixed.acidity         0.921  -0.158   0.229           0.262
volatile.acidity     -0.139          -0.629
citric.acid           0.500           0.764           0.143
residual.sugar                0.153   0.105           0.533
chlorides                             0.229  -0.292
free.sulfur.dioxide           0.670
total.sulfur.dioxide          0.986          -0.106   0.102
density               0.501                  -0.468   0.722
pH                   -0.715          -0.280   0.181   0.136
sulphates                             0.381           0.163
alcohol              -0.136  -0.107   0.254   0.949

                Factor1 Factor2 Factor3 Factor4 Factor5
SS loadings       1.909   1.494   1.384   1.253   0.971
Proportion Var    0.174   0.136   0.126   0.114   0.088
Cumulative Var    0.174   0.309   0.435   0.549   0.637

Test of the hypothesis that 5 factors are sufficient.
The chi square statistic is 795.53 on 10 degrees of freedom.
The p-value is 1.89e-164
```

Figure 3: Hypothesis testing

Therefore we can conclude that 4 factors are sufficient.

Based on the obtained factor loadings, we can discern the relationships between the physicochemical attributes and the extracted factors. Here's a breakdown of the loadings.

- **Factor 1 :** This factor appears to be positively associated with attributes such as fixed acidity, citric acid, free sulfur and density, while negatively associated with volatile acidity, alcohol and pH. This factor seems to capture aspects related to fixed acidity and PH value.
- **Factor 2 :** This factor is strongly negatively associated with volatile acidity. It is positively associated with free sulfur dioxide. Factor 2 may represent characteristics related to sulfur dioxide content.
- **Factor 3 :** Factor 3 shows a strong positive association with citric acid and residual sugar. It is negatively associated with density.This factor might represent characteristics volatile acidity, citric acid and sulphates content.
- **Factor 4 :** Factor 4 is positively associated with fixed acidity, chlorides, and density. It is weakly associated with pH. This factor seems to capture aspects related to alcohol and chloride content.

- **Factor 5 :** Factor 5 shows a strong positive association with alcohol. It is moderately associated with fixed acidity and density. Factor 5 may represent characteristics related to residual sugar content and density.

Communalities represent the proportion of variance in each variable that is accounted for by the factors extracted through factor analysis. Here's an interpretation based on the communalities obtained for each variable:

```
      fixed.acidity    volatile.acidity         citric.acid      residual.sugar           chlorides
          0.9950214           0.4330218           0.8540391           0.3201666           0.1434312
free.sulfur.dioxide total.sulfur.dioxide             density                  pH           sulphates
          0.4598086           0.9950009           0.9950198           0.6436915           0.1760047
            alcohol
          0.9950052
```

Figure 4: Communalities

- **Fixed Acidity:** 99.50% of the variance in fixed acidity is explained by the extracted factors. This suggests a strong association between fixed acidity and the underlying dimensions captured by the factors.
- **Volatile Acidity:** 43.30% of the variance in volatile acidity is explained by the extracted factors
- **Citric Acid:** 85.40% of the proportion of the variance in citric acid is accounted for by the extracted factors.
- **Residual Sugar:** 32.01% of the proportion of the variance in residual sugar is accounted for by the extracted factors.
- **Chloride:** 14.34% of the proportion of the variance in chloride is accounted for by the extracted factors.
- **Free Sulfer Dioxide:** 45.98% of the proportion of the variance in free sulfer dioxide is accounted for by the extracted factors.
- **Total Sulfer Dioxide:** 99.50% of the proportion of the variance in total sulfer dioxide: is accounted for by the extracted factors.
- **Density:** 99.51% of the proportion of variance suggests that nearly all of the variance in density is accounted for by the extracted factors. This indicates a very strong association between density and the underlying dimensions captured by the factors.
- **PH:** 64.36% of the proportion of the variance in PH is accounted for by the extracted factors.This suggests a significant association between pH and the underlying dimensions captured by the factors.
- **Sulphates:** 17.60% of the proportion of the variance in sulphates is accounted for by the extracted factors.

- **Alcohol:** 99.50% of the proportion of the variance in alcohol is accounted for by the extracted factors. This indicates a very strong association between alcohol content and the underlying dimensions captured by the factors.

3.2 Confirmatory Factor Analysis (CFA) :

Confirmatory Factor Analysis (CFA) is a statistical technique used to test the validity of a hypothesized factor structure by assessing how well the observed variables correspond to the latent factors.

Overall, these results suggest that the user model does not adequately fit the data, as indicated by poor fit indices such as CFI, TLI, and RMSEA. It may be necessary to revise the model or explore alternative factor structures to achieve better fit with the observed data.

```
lavaan 0.6.17 ended normally after 253 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        32

  Number of observations                          1599

Model Test User Model:

  Test statistic                              1532.688
  Degrees of freedom                                34
  P-value (Chi-square)                           0.000

Model Test Baseline Model:

  Test statistic                              8045.239
  Degrees of freedom                                55
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.812
  Tucker-Lewis Index (TLI)                       0.697

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)             -21695.933
  Loglikelihood unrestricted model (H1)     -20929.589
```

Figure 5: CFA1

4 Conclusion and recommendation:

From the Exploratory Data Analysis (EDA), it is evident that the physicochemical attributes of red wine exhibit complex interrelationships. The Factor Analysis revealed distinct latent factors underlying these attributes, with factors such as acidity, sulfur content, and alcohol concentration playing significant roles in determining wine quality. However, the Confirmatory Factor Analysis (CFA) results indicated that the proposed model did not adequately fit the observed data. This suggests that the hypothesized factor structure may not fully capture the underlying relationships among the variables.

While the EDA provided valuable insights, the limitations of the CFA underscore the need for further refinement or exploration of alternative factor structures. Future research should focus on incorporating additional variables beyond physicochemical attributes, such as sensory evaluations or wine making techniques, to provide a more comprehensive understanding of red wine quality.

5 References

References

Cortez, Cerdeira A. Almeida F. Matos T., Paulo, and J. Reis. UCI Machine Learning Repository.

6 Appendices

6.1 Dataset :

fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
7.4	0.700	0.00	1.90	0.076	11	34	0.99780	3.51	0.56	9.4	5
7.8	0.880	0.00	2.60	0.098	25	67	0.99680	3.20	0.68	9.8	5
7.8	0.760	0.04	2.30	0.092	15	54	0.99700	3.26	0.65	9.8	5
11.2	0.280	0.56	1.90	0.075	17	60	0.99800	3.16	0.58	9.8	6
7.4	0.700	0.00	1.90	0.076	11	34	0.99780	3.51	0.56	9.4	5
7.4	0.660	0.00	1.80	0.075	13	40	0.99780	3.51	0.56	9.4	5
7.9	0.600	0.06	1.60	0.069	15	59	0.99640	3.30	0.46	9.4	5
7.3	0.650	0.00	1.20	0.065	15	21	0.99460	3.39	0.47	10.0	7
7.8	0.580	0.02	2.00	0.073	9	18	0.99680	3.36	0.57	9.5	7
7.5	0.500	0.36	6.10	0.071	17	102	0.99780	3.35	0.80	10.5	5
6.7	0.580	0.08	1.80	0.097	15	65	0.99590	3.28	0.54	9.2	5
7.5	0.500	0.36	6.10	0.071	17	102	0.99780	3.35	0.80	10.5	5
5.6	0.615	0.00	1.60	0.089	16	59	0.99430	3.58	0.52	9.9	5
7.8	0.610	0.29	1.60	0.114	9	29	0.99740	3.26	1.56	9.1	5
8.9	0.620	0.18	3.80	0.176	52	145	0.99860	3.16	0.88	9.2	5
8.9	0.620	0.19	3.90	0.170	51	148	0.99860	3.17	0.93	9.2	5
8.5	0.280	0.56	1.80	0.092	35	103	0.99690	3.30	0.75	10.5	7
8.1	0.560	0.28	1.70	0.368	16	56	0.99680	3.11	1.28	9.3	5
7.4	0.590	0.08	4.40	0.086	6	29	0.99740	3.38	0.50	9.0	4
7.9	0.320	0.51	1.80	0.341	17	56	0.99690	3.04	1.08	9.2	6
8.9	0.220	0.48	1.80	0.077	29	60	0.99680	3.39	0.53	9.4	6

Figure 6: Dataset

The dataset can be find from this link.

6.2 R Codes:

Load Libraries$library(tidyverse)$
$library(corrplot)$
$library(ggplot2)$
$library(lavaan)$
$library(psych)$
$library(data.table)$
$library(factoextra)$
$library(knitr)$
$library(corrr)$
$library(ggcorrplot)$
$library(gridExtra)$
Load the dataset
$wine\_data <- read.csv("../data/winequality - red.csv", sep = ";")$
$head(wine\_data)$
$view(wine\_data)$
Standarize the data
$wine\_data <- apply(wine\_data, 2, scale)$
Principle component analysis
$pca\_results <- prcomp(wine\_data[, C(1:11)], scale. = T)$
$summary(pca\_results)$
$eigenVal <- get\_eigenvalue(pca\_results)$
$eigenVal$
Kaiser's Criterion
$num\_factors <- sum(eigenVal\$eigenvalue > 1)$

$num\_factors$
Scree plot
$fviz\_eig(pca\_results, addlabels = TRUE)$
$fa.parallel(wine\_data[, C(1:11)], fm = "pa", fa = "fa")$
Graph of variables
$fviz\_pca\_var(pca\_results, col.var = "red")$
Contribution of each variable
$fviz\_cos2(pca\_results, choice = "var", axes = 1:3)$
Factor Analysis For PC method
H_0 : 5 factors are sufficient vs H_1 : More factors needed
$fa\_results\_pc < -factanal(wine\_data[, C(1:11)], factors = 5, method = "principal", rotation = "varimax")$
$fa\_results\_pc$
factor loadings
$loadings\_pc < -fa\_results\_pc\$loadings$
$loadings\_pc$
Communality
$communalities\_pc < -rowSums(loadings\_pc^2)$
$communalities\_pc$
Factor Analysis For ML method
H_0 : 5 factors are sufficient vs H_1 : More factors needed
$fa\_results\_ml < -factanal(wine\_data[, c(1:11)], factors = 5, method = "ml", rotation = "varimax")$
$fa\_results\_ml$
factor loadings
$loadings\_ml < -fa\_results\_ml\$loadings$
$loadings\_ml$
Communality
$communalities\_ml < -rowSums(loadings\_ml^2)$
$communalities\_ml$
Confirmatory Factor Analysis (CFA)
$cfa\_model < -'$
$Factor1 = \ fixed.acidity + pH$
$Factor2 = \ free.sulfur.dioxide + total.sulfur.dioxide$
$Factor3 = \ volatile.acidity + citric.acid + sulphates$
$Factor4 = \ chlorides + alcohol$
$Factor5 = \ residual.sugar + density$

$fixed.acidity \ \ fixed.acidity$
$volatile.acidity \ \ volatile.acidity$
$citric.acid \ \ citric.acid$
$residual.sugar \ \ residual.sugar$
$chlorides \ \ chlorides$
$free.sulfur.dioxide \ \ free.sulfur.dioxide$
$total.sulfur.dioxide \ \ total.sulfur.dioxide$
$density \ \ density$

$pH \sim pH$
$sulphates \sim sulphates$
$alcohol \sim alcohol$
$'$

Fit the CFA model

$cfa\_fit <- sem(cfa\_model, data = wine\_data[, c(1:11)])$

Summarize the model fit

$summary(cfa\_fit, fit.measures = TRUE)$