**MAR 6669: Assignment 2**

Due date: Thursday, April 17, 11:59 PM EST. Upload submissions to Canvas.

Organize all figures, numeric results, discussions, and conclusions in a single PDF. Each group should submit **one PDF report**, and related code scripts.

# Wine qualities (7 points)

The data in wine.csv contains information on 11 chemical properties of 6500 different bottles of *vinho verde* wine from northern Portugal. In addition, two other variables about each wine are recorded:

whether the wine is red or white the quality of the wine, as judged on a 1-10 scale by a panel of certified wine snobs.

Run PCA algorithm on the 11 chemical properties and then run K-means **on the principle components** you choose. Summarize your results. (Set the number of clusters K=2 for this question)

We would like to explore whether the chemical properties can distinguish the red wine and white wine. Compute the number of white and red wines in each cluster. Does the proportion of red differ in the two clusters?

First, we run K-Means on Principle Components Analysis. We decided to run on 3 components based on the elbow chart and the eigenvalues. As you can see in the chart below, at component =3, the cumulative explained variance is at 0.64, which is large enough and hits the threshold. And based on the graph, the elbow occurs at component = 3.

A graph with a line

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

Then, we perform PCA and gets how many red and white wines are in each cluster.

A computer code with text

AI-generated content may be incorrect.

A computer screen shot of a computer code

AI-generated content may be incorrect.

A computer screen shot of a computer code

AI-generated content may be incorrect.

Based on the results, we can see that the proportion of red wine differs among the 2 clusters. 85% of red wine is in cluster 0 and 15% of red wine is in cluster 1. Red wines are mostly grouped into Cluster 0. White wines are more balanced but tend slightly toward Cluster 1. And 39.5% of the samples in cluster 0 is red wine and 7.9% of the samples in cluster 1 is white wine.

Bonus question: Does your unsupervised technique also seem capable of distinguishing the higher from the lower-quality wines? This can be assessed using hist plots of wine quality for each cluster.

A graph of different sizes and numbers

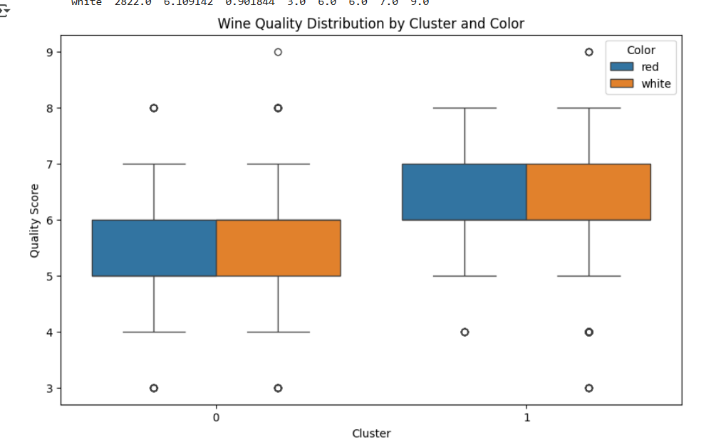
AI-generated content may be incorrect.

Based on the graph shown above, we can clearly see the unsupervised technique seems capable of distinguishing the higher quality wine from the lower quality wine. In cluster 1, most samples have quality scores of 5 and 6. While in cluster 2, most samples have quality score of 6 to 7.

**Data visualization**

**A number of numbers on a white background

AI-generated content may be incorrect.**



* Based on the boxplot, we can see that cluster 1 has higher quality score than cluster 0. There is no difference in quality score across wine color.

A screenshot of a graph

AI-generated content may be incorrect.

Left plot: red wine

Points are spread out more vertically, suggesting more variation along Principal Component 2.

There is no strong separation between quality scores (colors are mixed), but:

* Higher quality (redder) scores seem more concentrated around the middle-lower region.
* Lower quality (bluer) scores are somewhat more dispersed but still not clearly separated.

Right plot: White wine

The point cloud is more elongated horizontally — variation is mostly along Principal Component1.

There seems to be a slight gradient:

* Lower quality wines (blue) are more concentrated to the left side.
* Higher quality wines (red) trend slightly rightward.
* No clear-cut cluster by quality

A diagram of the different types of substances

AI-generated content may be incorrect.

* Since there is a 90-degree angle between density and sulfur dioxide, we can tell that there is no correlation between these two variables
* Alcohol is negatively related to density
* Free sulfur dioxide and total sulfur dioxide are strongly positively correlated.

To clarify: I'm not asking you to run a predictive algorithm to predict wine type. Rather, I'm asking you to see whether the differences in the labels (red/white and quality score) emerge naturally from applying an unsupervised technique to the chemical properties.

**Hint**: remember to standardize data before using PCA.

# Car market (8 points)

The file car\_data.csv contains consumers' responses to questions about attitudes toward buying a new car. The details of these questions can be found in the data key, car\_data\_key.docx.

We are interested in using these psychographic measures to understand the new car market, and segment customers into different buyer types. **For the following questions, you should ignore the Purchased variable.**

1. Using k-means clustering on the psychographic questions (Q1-Q17) and the Ideal\_Pricevariable, identify and justify an appropriate number of attitudinal segments (less than 6) for new car buyers (Hint: use the within-cluster sum of squares).

A graph with a blue line

AI-generated content may be incorrect.

* Based on the elbow chart, we can see the elbow occurs at k= 3 or k=4.
* Then we run a silhouette analysis on which k is better.

A close-up of a screen

AI-generated content may be incorrect.

* Based on the silhouette analysis, k=4 is a better choice because it scores higher (0.47) than k =3 (0.407)

1. Describe each of the above segments by computing how many respondents are in each segment, and by characterizing the mean of the question responses for each segments.

A screenshot of a computer code

AI-generated content may be incorrect.A screenshot of a paper

AI-generated content may be incorrect.



1. Based on the computations from part b, name each segment, and create a short, 1-3 sentence, descriptive profile of each segment.

Using these descriptions, we can refine the profiles for each segment:

1. **Segment 0: Trend Enthusiasts**
   * **Profile**: This segment shows generally positive responses across most questions, with particularly high scores in Q9 to Q12. They seem to value trendiness, fashion, and making a statement with their car.
2. **Segment 1: Discontented Respondents**
   * **Profile**: This segment consistently scores low across all questions, indicating a generally negative perception. They are less satisfied with all aspects of car features and characteristics.
3. **Segment 2: Price-Sensitive Critics**
   * **Profile**: This segment is highly sensitive to price (Ideal\_Price) and has very low scores on most questions, except for a slight positive on Q9 to Q12. They are critical and price-conscious but still care about trendiness and fashion to some extent.
4. **Segment 3: Enthusiastic Supporters**
   * **Profile**: This segment has very high scores on Q1 to Q8, indicating strong enthusiasm for features, quality, performance, and driving experience. However, they show some dissatisfaction in Q9 to Q12, suggesting they are less concerned with trendiness and fashion.
5. Ignoring, for the moment, the segments you just described, use Factor Analysis to identify and justify an appropriate number of factors (at least 2) describing consumers’ responses to the psychographic questions (this time **NOT** including the Ideal\_Price variable). You can use the number factors as 3 for this question. Report the (varimax rotated) factor loading matrix.

A close-up of numbers

AI-generated content may be incorrect.

A graph with a line graph

AI-generated content may be incorrect.

* Based on the explained variance and the elbow chart, 3 is a good number for factors

A screenshot of a graph

AI-generated content may be incorrect.A table of numbers with numbers

AI-generated content may be incorrect.

* Based on the factor loading, you can see that factor 1 loaded heavily on questions 1-8 and factor 2 loaded heavily on questions 9-14 and factor 3 loaded heavily on questions 15-17

Bonus question: Name and interpret the factors you found in part d. Substantively, what are the factors capturing?

A screenshot of a computer

AI-generated content may be incorrect.

We named factor 1 as car features and driving experience, and we named factor 2 as car style and emotional expression and we named factor 3 as emotional/cultural attachment. Based on the survey, questions 1-8 focus on performance and functional appeal. Question 9-14 focus on style and image orientation. Question 15-17 focus on cultural and emotional attachment