Practice (HW) 2

Due: February 14, 2019

- Asking questions to TAs and collaborating with classmates are encouraged, but copying, sharing, or distributing any material is strictly prohibited. Homework should be students' original work.
- Please submit
  1) SAS code (.SAS) with detailed comments
  2) PDF document with relevant output and interpretations
- Late homework will not be accepted.

Wine Quality

A group of researchers collected 1599 red and 4898 white wine samples. Each sample was evaluated by at least three wine experts, who graded the wine in a scale that ranges from 0 (very bad) to 10 (excellent). The final sensory score is given by the median of these evaluations. For each wine, following information was gathered:

Variable	Description
facid	Fixed acidity (g(tartaric acid)/dm3)
vacid	Volatile acidity (g(acetic acid)/dm3)
cacid	Citric acid (g/dm3)
rsugar	Residual sugar (g/dm3)
chlorides	Chlorides (g(sodium chloride)/dm3)
freesd	Free sulfur dioxide (mg/dm3)
totalsd	Total sulfur dioxide (mg/dm3)
density	Density (g/cm3)
pH	pH values
sulphates	Sulphates (g(potassium sulphate)/dm3)
alcohol	Alcohol (vol.%)
quality	Quality of wine (0: very bad – 10: excellent)

This dataset is publicly available for research. The details are described in
Cortez, Paulo, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. "Modeling wine preferences by data mining from physicochemical properties." *Decision Support Systems* 47, no. 4 (2009): 547-553.

a) Import the dataset: The dataset ‘wine.xlsx’ has two sheets: ‘white’ and ‘red’. Use PROC IMPORT to import both sheets and name them ‘white’ and ‘red’, respectively. Define a new variable ‘color’, which indicates the color of each wine. Combine (Stack) the two datasets and name the merged dataset ‘wine’.

b) Label the variables as follows:

Variable	Label
facid	Fixed acidity
vacid	Volatile acidity
cacid	Citric acid
rsugar	Residual sugar
freesd	Free sulfur dioxide
totalsd	Total sulfur dioxide

Print the first 5 observations of the dataset with labels to check if labels were applied correctly.

c) Create a new variable called ‘qgroup’. If the sensory (quality) score of wine is less than 5, set qgroup = poor. If the score is 5 ≤ score < 7, then qgroup = normal. If the score is greater than or equal to 7, then set qgroup = great.

d) Descriptive statistics: Use the dataset ‘wine’ to provide the following tables & plots and <u>describe the distribution</u> (e.g. missing values, symmetry, skewness, association between variables, location (mean, median), dispersion (range, standard deviation), outliers) of variables displayed in those tables and plots.

d1) Free sulfur dioxide

i. For each wine color, calculate **minimum**, **maximum**, **mean**, and **median** of free sulfur dioxide. Use 2 decimal points to present those descriptive statistics.

ii. Plot **histograms** to visualize how the distribution of free sulfur dioxide is different depending on wine color. Use ‘count’ as the scale of vertical axis. Overlay the histogram with density curve with type=kernel. (Hint: Panel)

iii. Create a **binary variable** ‘fsd’ indicating whether or not the wine contains less than 5 mg/dm3 free sulfur dioxide. Produce a **two-way frequency table** (norow nopercent) between qgroup (column) and the new variable fsd (row).

d2) Quality of wine

i. Produce **summary statistics** (min, mean, std, median, max) of sensory (quality) score. Use two decimal points.
ii. Produce a **frequency table** of score qgroup.
iii. Provide an appropriate **plot** for describing the distribution of qgroup.
iv. Create **boxplots** of sensory (quality) score to see if there is a difference in distribution of sensory (quality) score between red and white wine.
v. In order to check the relationship between the quality of wine and the free sulfur dioxide, create a **scatterplot** with the free sulfur dioxide on x-axis and sensory (quality) score on y-axis. Use different colors for wine color. Describe what you learn in the plot.

d3) Alcohol

i. Use any appropriate (graphical / tabular / numeric) tool to check the **distribution** of alcohol.
ii. For each wine color, check if the distribution of alcohol is **normal**. Provide both numerical and graphical evidences to draw a conclusion.