Homework 3 (Due at 6:30pm, Sep 26)

Problem 1: Download the wine data set at UCI:

https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data

Dataset description: https://archive.ics.uci.edu/ml/datasets/Wine

1) There are 14 variables in this dataset. INPUT the 14 variables by first four letters for each variable, e.g. ‘alco’ for ‘Alcohol’. Then use LABEL statement to name the full name for each variable. NOTES: The first variable is ‘identifier’ not ‘Alcohol’.
2) Change the dataset’s TITLE to ‘Wine Summary Data’.
3) Generate descriptive statistics by using PROC MEANS.
4) Generate a probability plot and test the normality for variable ‘Malic acid’. (i.e normal or not)
5) Generate a vertical histogram for ‘Malic acid’ divided by variable ‘identifier’ (the first variable, use subgroup option)
6) Generate a scatter plot for ‘Malic acid’ vs ‘Ash’ for each value of ‘identifier’. (Use BY statement).

Submit code for parts 1) to 6), output 3) to 6).

Problem 2: Download the corn data which measurements in six states from 1909 to 1927 at sakai, also see dataset description from the file corn_info from sakai.

Create a new variable catyield to categorize the corn yield:
if yield < 32, then catyield = ‘poor’;
if yield >= 32, then catyield = ‘good’.

Create a new variable catrain to categorize the amount of rainfall:
if rain < = 9.7, then catrain = ‘drought’;
if 9.7 < rain < = 12, then catrain = ‘normal’;
if rain >12, then catrain = ‘wet’.
(Hint: please see the class 2 notes on ‘create new variables’ if needed.)

1) Use PROC FREQ to analyze the variables catyield and catrain. Explain the SAS output.

____

____

____

____

2) Create a two way table for corn yield (i.e., catyield) by the amount of rainfall (i.e., catrain). Do you notice anything unusual? Why? (i.e. how does the amount of rainfall relates to the corn yield?)

____

____

____

____

Please hand in your SAS code and output and provide your answers to the above questions.