**Problem 1 (25% points):**

Answer each of the following questions with a few sentences:

1. What is the difference between *classification* and *clustering* (5%)?
   1. Clustering is more of an exploratory process than classification. In clustering you do not have prediction variables and you cannot calculate and error of prediction. In classification you have a data set with predictor values and you are able to calculate a prediction and error.
2. In a data table, what do the *columns* represent and what do the *rows* represent (5%)?
   1. The columns = attributes
   2. The Rows = records

c. After loading new data into SPSS, describe two tasks you might do to clean your data

(5%).

1. Fill in missing values.

2. Identify duplicate cases

d. Explain which type of data mining algorithm (also called data mining functionality)

would you use to answer each of these questions and why?

1. What are five groups of customers who buy similar things (5%)?
   1. Clustering.

ii. I sell milk – can I predict if a user will buy that based on the other things they

bought (5%)?

Association and correlation analysis

**Problem 2 (25% points):**

Explain in few words whether or not each of the following activities is a data mining task and why.

1. Dividing the customers of a company according to their gender (5%).
   1. This would be considered a clustering data mining task because you are using the data to create different groups based on certain variables.
2. Computing the total sales of a company (5%).
   1. This is not a data mining task. This is a math calculation. We are not making a decision based on the data.
3. Sorting a student database based on student identification numbers (5%).
   1. This is not a data mining task.
4. Estimating the probability of the outcomes of tossing a (fair) pair of dice (5%).
   1. This is not a data mining task; this is a statistical probability analysis.
5. Predicting the future stock price of a company using historical records (5%).
   1. This is a data mining task because we are using historical records to train an algorithm to predict the future stock. We are using the data to make the prediction.

**Problem 3**

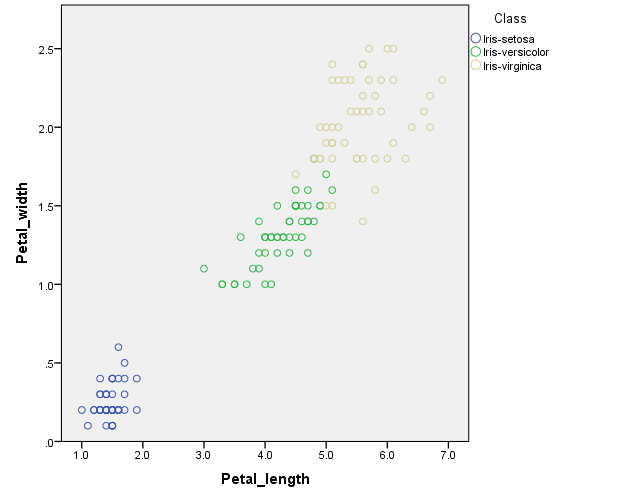
a. Visualize the relationship between the two sepal variables, sepal length and sepal width, using a scatter plot. Use different colors or symbols per class so you can see how the classes are related to this pair of variables. We talked in class about how classifiers work, broadly-speaking. Do you think that a classification algorithm using these two variables will be successful in classifying data with respect to the class labels we have? Explain why or why not and include the plot image with your answer (10%).

No because the Iris-versicolor and Iris-virginica are not distinguishable in the graph. They have similar lengths and widths, making it hard to group them into two separate classes based on width and length.



b. Repeat part (a) for the petal variables (10%).

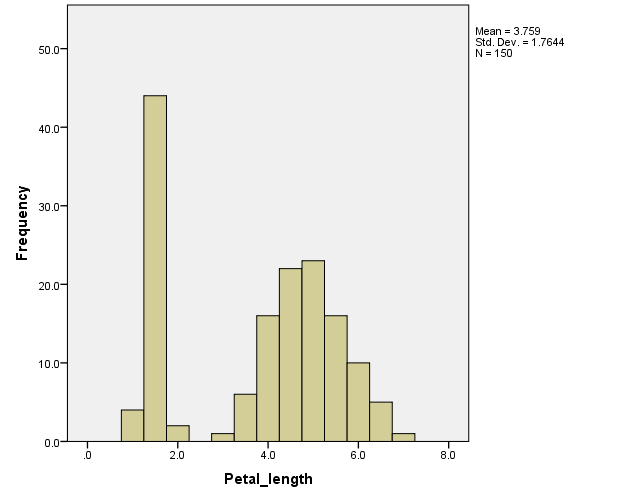
Based on petal\_width and petal\_length you can classify the data. You can see the three distinct classes are separated in the graph.



c. Create a histogram for each of the four variables. Histograms in SPSS are just a different graph type from scatterplots. Describe what you can tell about the distribution of each variable (10%).

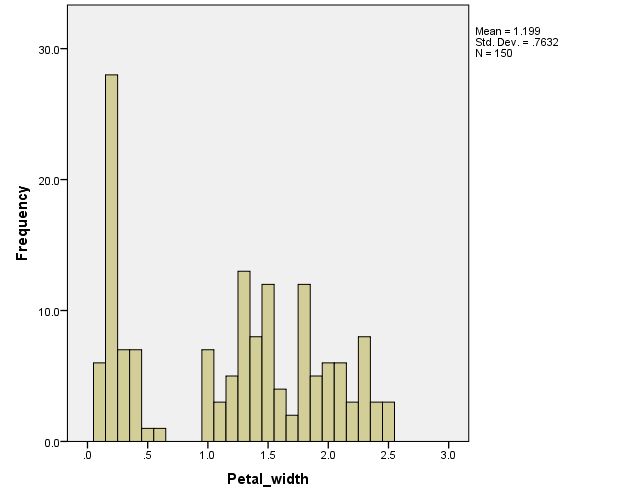
Petal\_Length:

This is a bimodal distribution. The distribution shows that there are two peaks at about 1.5 and 5. Looking at the mean alone (3.7) is not a good representation of the data.



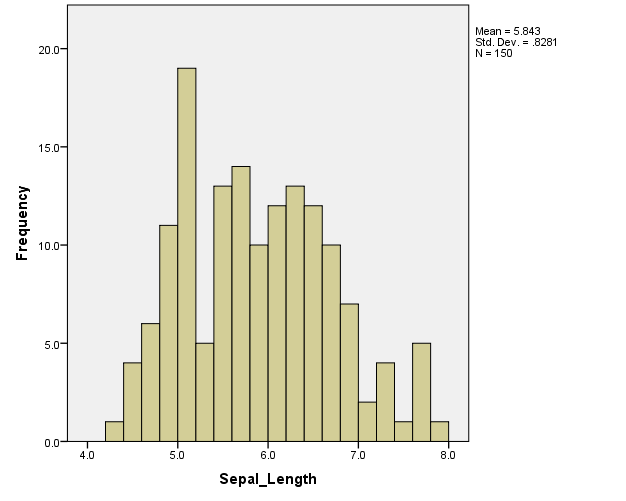
Petal\_Width:

The petal\_width distribution is not normal. IT appears that you could group the data into at least two groups since there is a large frequency of petal\_width from 0 to .5 and the rest between 1-2.5



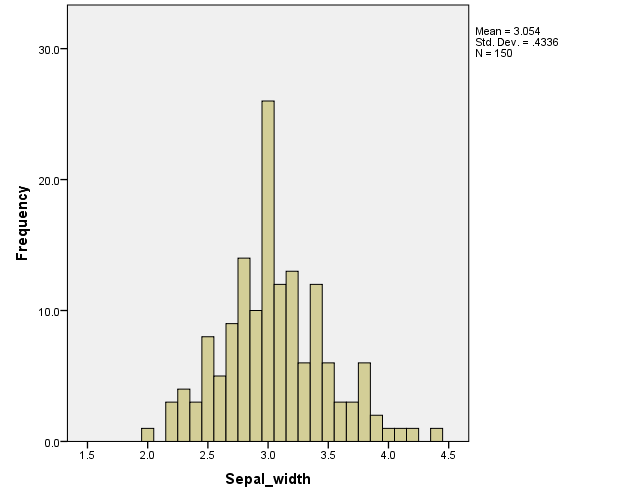
Sepal\_length:

The sepal length distribution is more normal than the petal distributions. There is a slight skew right due to the high frequency of sepal\_length around 5.0.



Sepal\_width:

The depal\_width distribution is the most normal distribution of each variable. You can see that the mean is 3.054, which is there the highest frequency of the data is (shown by the peak in the center of the graph).



d. Determine if there are any outliers in the data with respect to the sepal length (10%).

There are no outliers with respect to sepal length, because there are no values with a |z-score| > 3

e. Repeat d. for the petal length (10%).

There are also no outliers with respect to petal length, because there are no values with a |z-score| > 3