IS 467
Homework #1

Name: Kai Chung Ying
Date: 04-16-2017
Email: kying@mail.depaul.edu

**Problem 1 (5 points):**

Answer each of the following questions with a few sentences:

a. What is the difference between *classification* and *clustering*?

Ans:

Classification is one kind of supervised operation assigns records (Training data) with pre-defined class to algorithm turns into a model or classifier to predict the Testing Data which Class/Label they might belong to. Classification application includes Decision Tree and k-nearest neighbor methods etc.

Clustering is one kind of unsurprised operation (Because No pre-defined Class) to use algorithem groups the records into different clusters based on their similarity. Real-World applications include Fraud detention and Medical insurance Fraud claim.

b. What is the difference between *data warehouse* and *database*?

Ans:

Database is designed to store relational data for day-to-day operation for business organization. One of the typical type is OLTP (Online Transaction Processing) Database.

Data Warehouse typically stores all types of historical data and time-variant data for analytic purpose. Data Warehouse is larger than typical database for storing relational records. Data warehouse The one kind of Data Warehouse is called OLAP (Onine Analytic Procesing) database. The data stores in data warehouse is never updated. The data record or queried result is only responded to the end-user (Decision Maker).

c. What is the difference between *data mining* and *OLAP*?

Ans:

OLAP (Online Analytical Processing) is basically summarize the existed set of data by using more complex or user friendly programing and then used the result to predict for the future value, for example companies analyze their sales record to predict their inventory and sales.

Data mining is a process to mine the data and discover the hidden pattern among the dataset which might be valuable for future business use, e.g. data mining could discover the customer purchase habits by analyzing the habitat of how they search the ecommerce store (e.g. Amazon.com or ebay.com)

d. What is the difference between *data marts* and *data warehouse*?

Ans:

As answer on the part b, Data Warehouse is used to store integrated data for analytic purpose. Basically Data Marts is used as the SAME function as Data Warehouse, but generally their size is smaller than Warehouse. D.W. is designed for multiple subject use, D.M is designed for Single subject use. For the above reason, D.W. is typically built for entire organization and D.M. is for smaller organization or subset of department for use.

Name: Kai Chung Ying
Date: 04-16-2017
Email: kying@mail.depaul.edu

e. In a data table, what do the *columns* represent and what do the *rows* represent?

Ans:
Column represent attribute or field.
Rows represent a tuple or record.

**Problem 2 (5 points):**

Answer each of the following questions with a few sentences:

a. (2 points) After loading new data into SPSS, describe two tasks you might do to clean your data.

Ans: Two Tasks: Removal of Outliers by checking how the data distributed and Filling in the Missing Values.

b. Explain which type of data mining algorithm (also called *data mining functionality*) would you use to answer each of these questions?

i. What are five groups of customers who buy similar things?

Ans: Clustering because we are going to group the customers based on the purchase similarity. There would be NO pre-defined class for this analysis.

ii. What are different sets of products that are often purchased together?

Ans: Clustering because we are going to group the products based on their purchased record from customers. There would be NO pre-defined class for this analysis.

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

Ans: Classification because we have pre-defined class (Buy or NOT buy). Result will be predicted by using provided variables (i.e. other things they bought).

**Problem 3 (5 points):**

Explain whether or not each of the following activities is a data mining task.
a. Dividing the customers of a company according to their gender.
b. Computing the total sales of a company.
c. Sorting a student database based on student identification numbers.
d. Predicting the outcomes of tossing a (fair) pair of dice.
e. Predicting the future stock price of a company using historical records.

Ans:

Based on the definition (PPT pg.22), Data Mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful and ultimately understandable patterns in data.

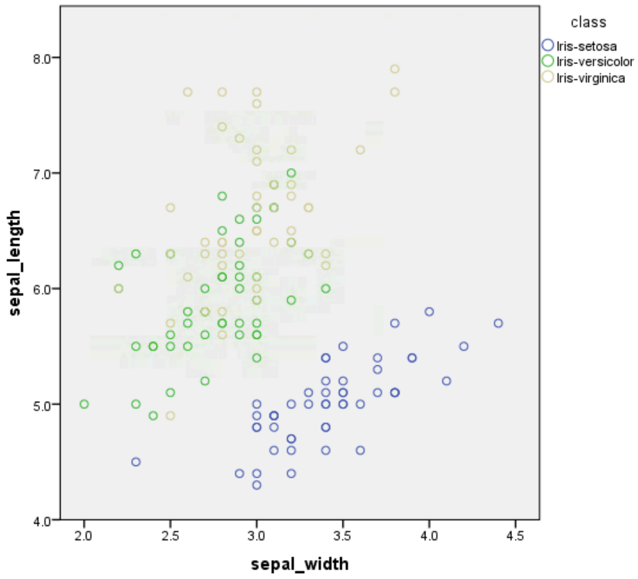
a. No, it is NOT a data mining task because this task is not satisfied any factors above. This task can be completed based on the record from the company. This task could be only sort out the records of table and divide the customers according to their gender. No further exploration is needed.
b. No, it is NOT a data mining task because this task is just mathematical calculation process. No Research or pattern exploration is required to complete this task
c. No, it is NOT a data mining task because it is just a sorting operation by using some type of operational programming language. No further pattern exploration is needed. The result might NOT be useful for future use.
d. Yes, it is a data mining task because this task requires collecting certain amount of data from some training data, and then explore the pattern of the data by Math or Statistics in order to generate some useful predictive and understandable result for future prediction.
e. Yes. It is a data mining because this task includes the above factors. By using the historical records to reveal the pattern of stock price. With this information, we should be able to predict some useful trends of the future stock price. And we properbly need to

**Problem 4:**

a) Visualize the relationship between the two sepal variables, sepal length and sepal width, using a scatter plot.

**GGraph**

```
[DataSet2] \\tsclient\home\Desktop\Depaul\IS467\wk2\is467_wk1_hw_04162017.sav
```



a.

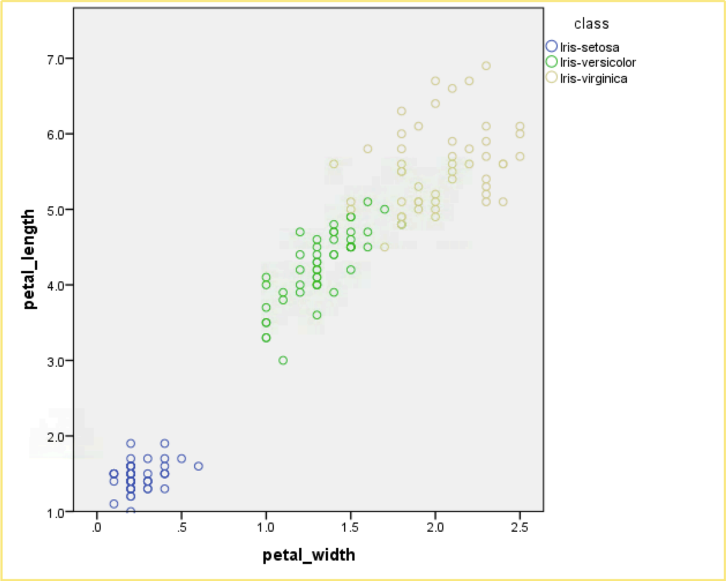
Use different colors or symbols per class so you can see how the classes are related to this pair of variables. We talked in class about how classifiers work, broadly-speaking. Do you think that a classification algorithm using these two variables will be successful in classifying data with respect to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

Ans:

By observing on the above graph, we possibly could use classification algorithm to PARTIALLY classify data. Iris-setosa (Blue color dots) are obviously categorize out from other 2 classes (Iris-versicolor, Iris-virginica). However, the other 2 classes (Iris-versicolor, Iris-virginica) are harder to classify by using sepal_width and sepal_length based on this scatter plot. Their data (Green and Light Green color) are more blended around the area.

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

**GGraph**

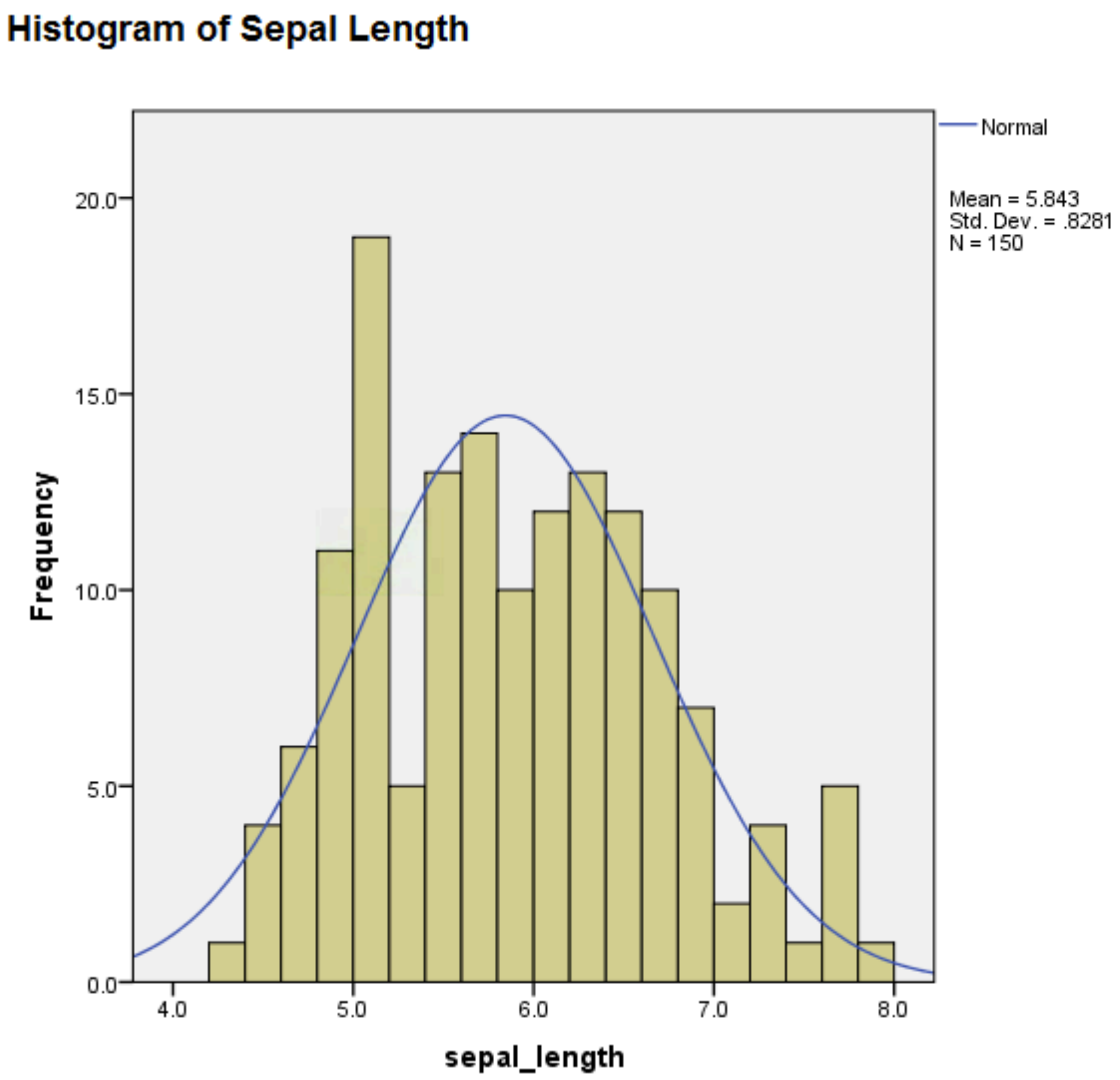


Use different colors or symbols per class so you can see how the classes are related to this pair of variables. We talked in class about how classifiers work, broadly-speaking. Do you think that a classification algorithm using these two variables will be successful in classifying data with respect to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

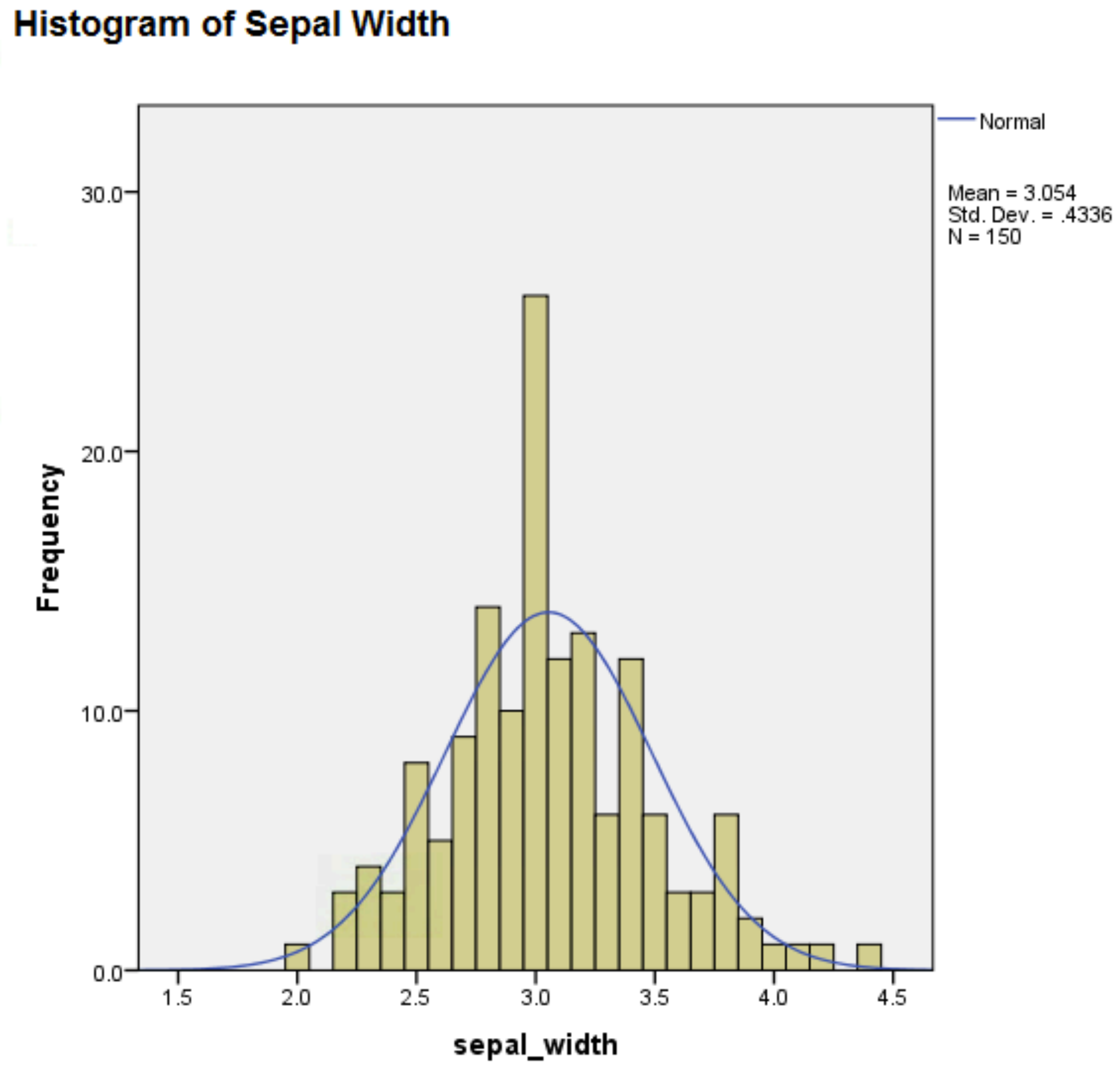
Ans:
By observing the above scatter plot, I think that we could use classification algorithm to classify the data because these 2 predictor variables (petal_width, petal_length) clearly categorizes out into 3 different classes area. After running the algorithm, the classifier could be used to classify the new coming data(Testing set).

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.

Ans: Sepal Length histogram shows data is very much Normally distributed with mean of 5.843 and Std Dev = 0.8281

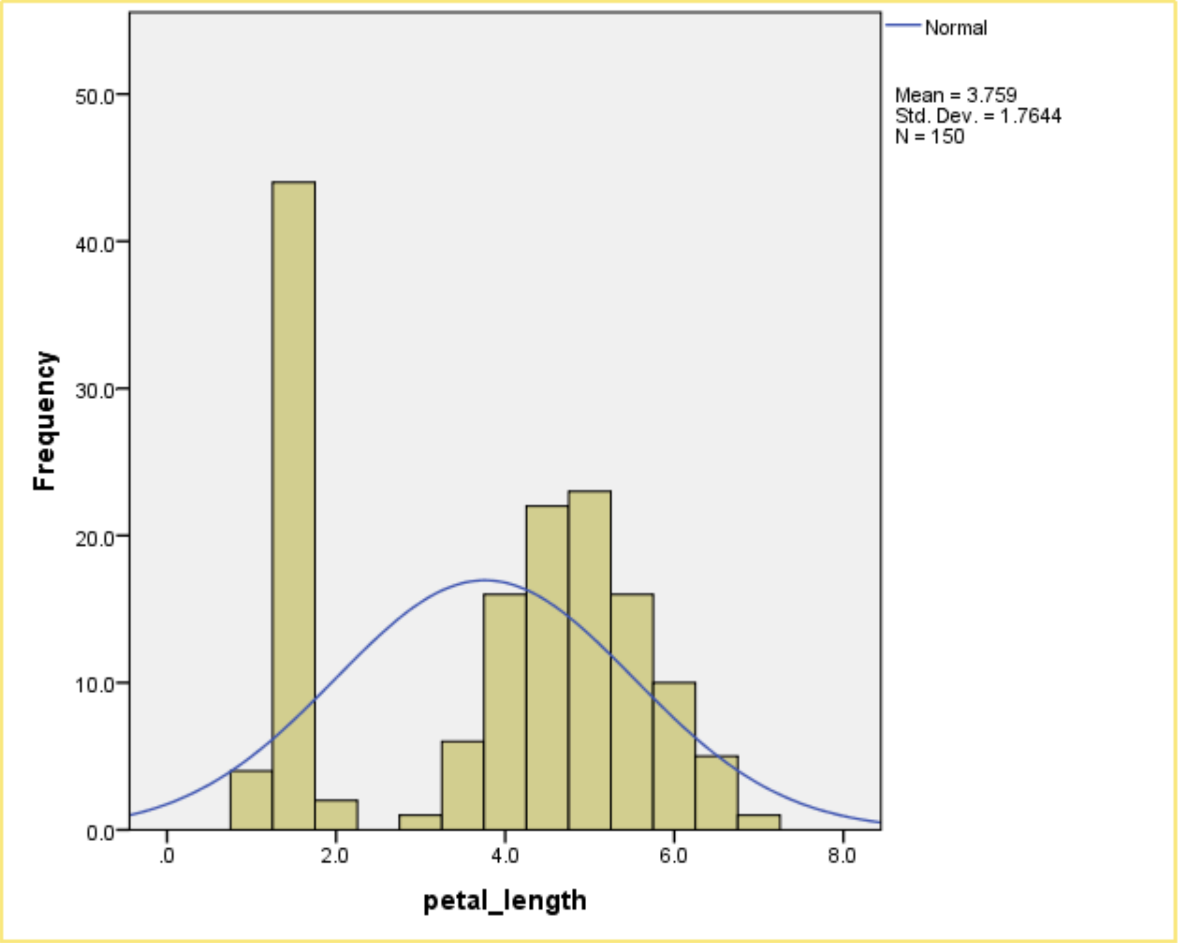


Ans: Sepal Width histogram shows data is very much Normally distributed with mean of 3.054 and S.D.=0.4336

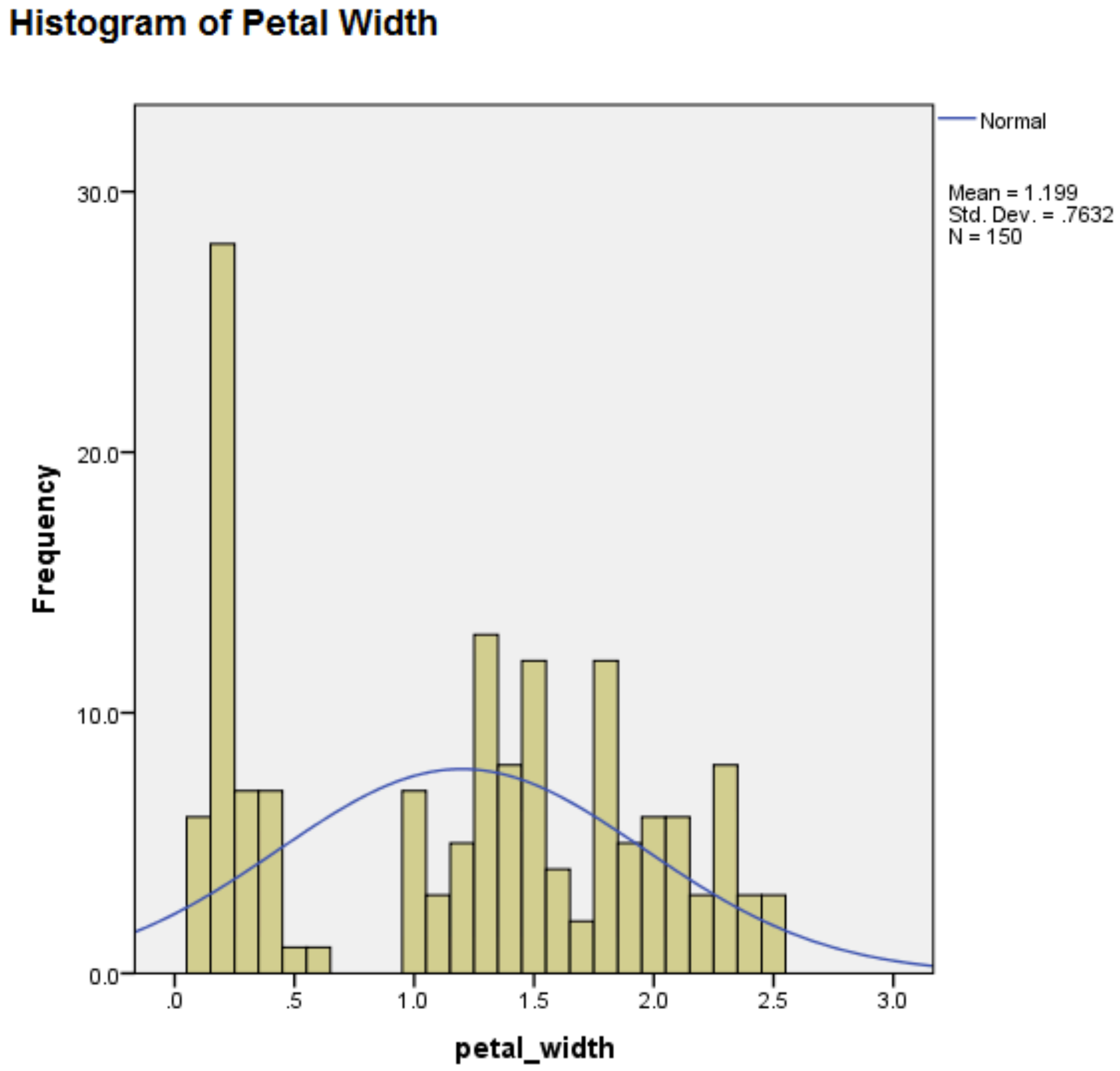


Ans: Petal Length Histogram shows Data is very much Normally distributed with mean 3.759 and Std. Dev 1.7644

**Histogram of Petal Length**

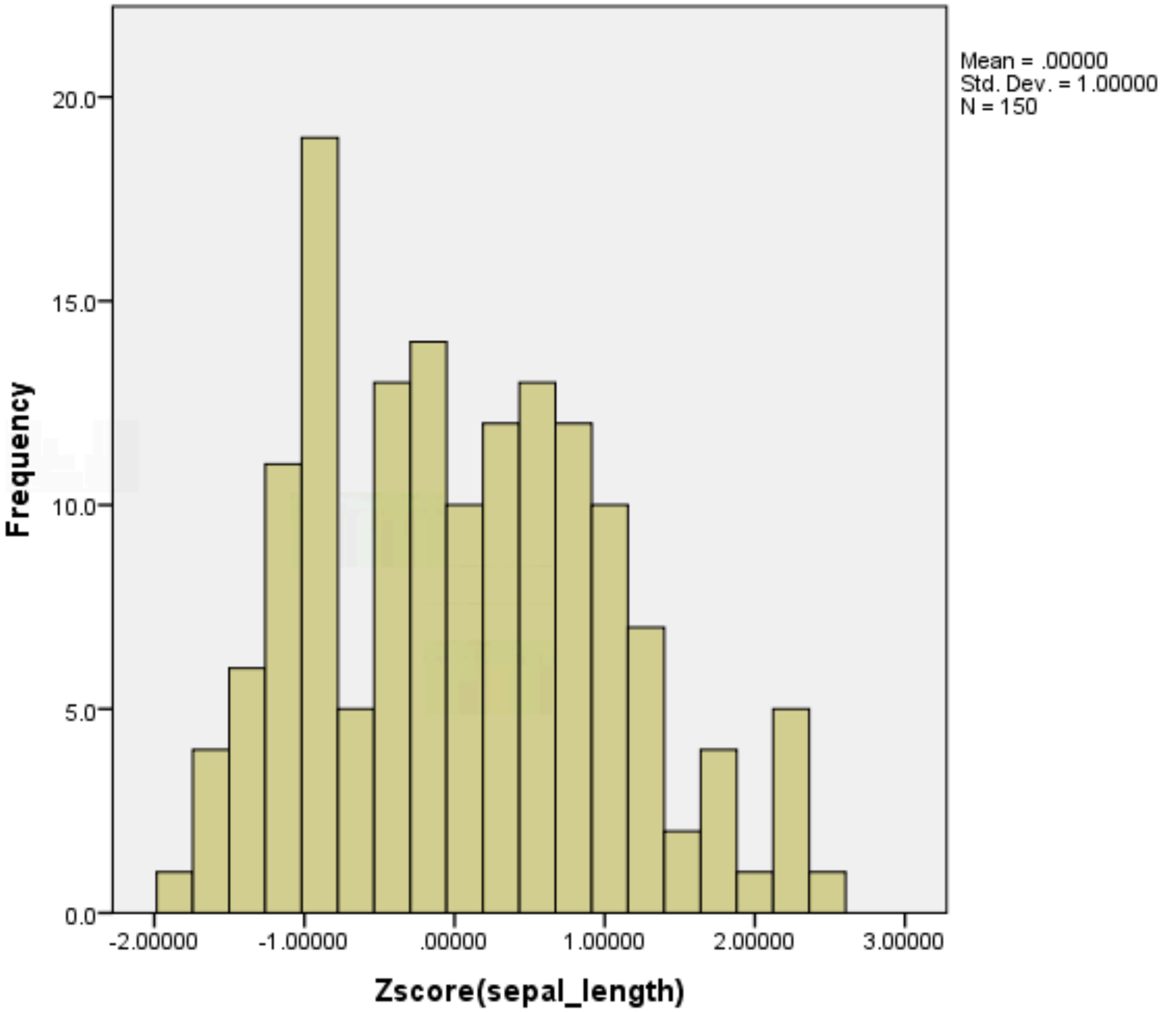


Ans: Petal Width Histogram shows Data is Right Skewed Distributed.



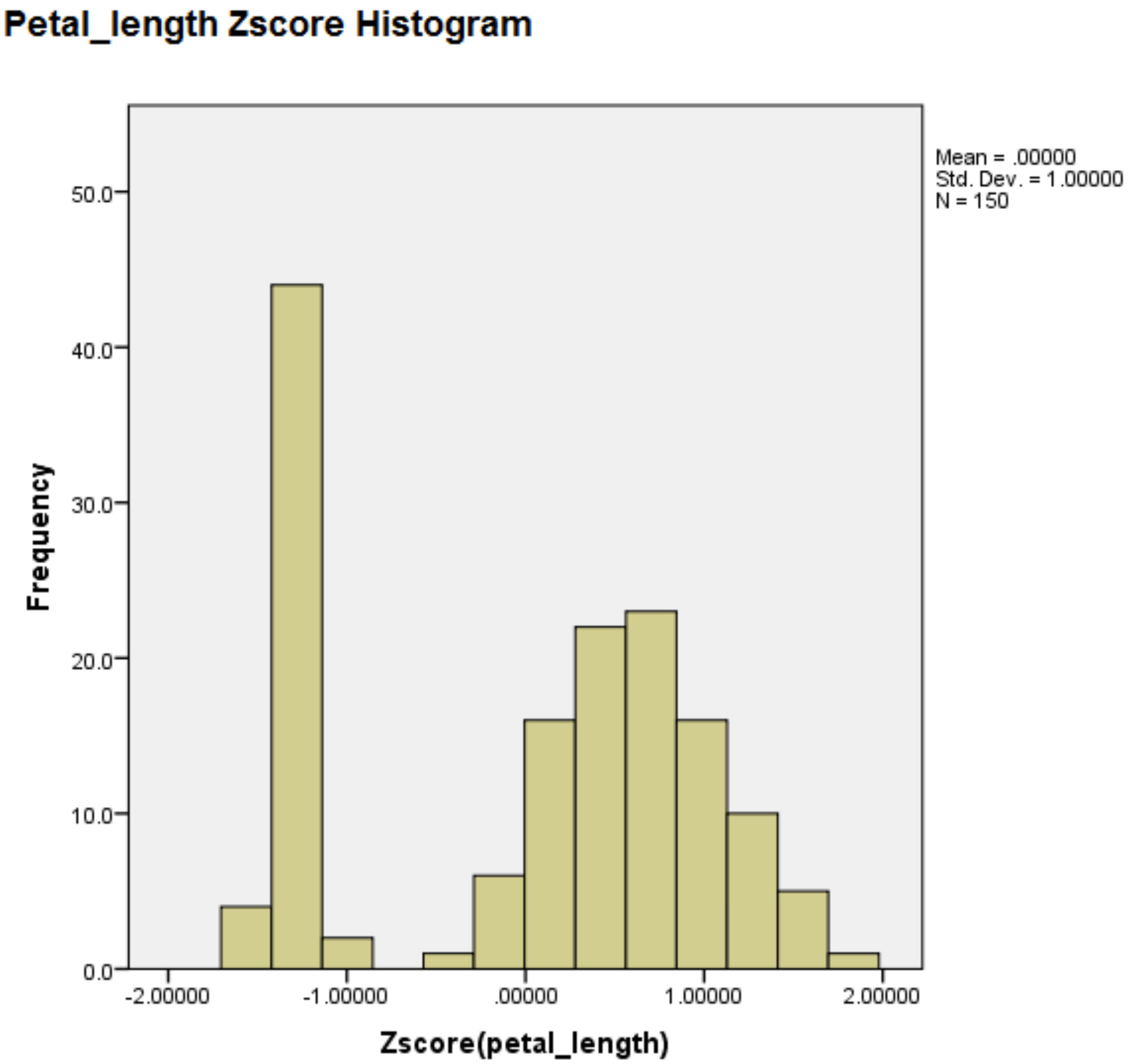
d) Determine if there are any outliers in the data with respect to the sepal length.

### sepal_length Zscore Histogram



No, I believe that there is NO outliers on the sepal length based on the above histogram because there is NO observations outside the overall dataset pattern

e) Repeat d. for the petal length.



No, I believe that there is NO outlier on the petal length based on the above histogram because there is NO observations outside the overall dataset pattern