1 **Obtain one of the data sets available at the UCI Machine Learning Repository and apply as many of the different visualization techniques described in the chapter as possible. The bibliographic notes and book Web site provide pointers to visualization software.**

Assignment 3 question 1.ipynb

2 **Identify at least two advantages and two disadvantages of using color to visually represent information.**

Advantages:

- Colors are helpful to speed up visual search: Color coding is a way to convey information quickly, which facilitates search. For example, Washington D.C. metro map.
- Colors are useful to transfer simple information to users and avoids the need to understand local language. For example, traffic lights colors are better than a sign in local language.

Disadvantages:

- Using colors to present information is an art. One must be very thoughtful about choosing the appropriate color scheme. Some time, bad color scheme delivers the wrong message.
- Colorblind people may not identify Green and Red accurately. Therefore, using color to visually represent information is challenging at times.

3. **What are the arrangement issues that arise with respect to three-dimensional plots?**

The key issue for three- dimensional plots arrangement is how to display the information so that as little information is obscured as possible. As the dimension of plots increases, the complexity of a visual representation of the data increases and it becomes harder for the intended audience to interpret the information.

4. **Discuss the advantages and disadvantages of using sampling to reduce the number of data objects that need to be displayed. Would simple random sampling (without replacement) be a good approach to sampling? Why or why not?**

Simple random sampling is an unbiased approach to gather the responses from a large group. Simple random sampling has inherent drawbacks. These disadvantages include the time needed to gather the full list of a specific population, the capital necessary to retrieve and contact that list, and the bias that could occur when the sample set is not large enough to adequately represent the full population.

6 **Describe one advantage and one disadvantage of a stem and leaf plot with respect standard histogram**.

Stem and leaf plots are a type of histogram, can be used to provide insights into the distribution of one-dimensional integer or continuous data.

On the other hand, a stem and leaf plot become rather cumbersome for large number of values.

8 **Describe how a box plot can give information about whether the value of an attribute symmetrically distributed. What can you say about the symmetry of the distributions of the attributes shown in Figure 3.11?**

Box plots are a useful method for showing the distribution of values of a single numerical attribute. The box plot has a box in the middle and whiskers on both sides of the box. If the line representing median of data is in the middle of the box, then the data is symmetrically distributed, at least in terms of 75% of the data between first and third quartiles. For the remaining data, length of whiskers and outliers is also an indication, although, since these features do not involve as many points, they may be misleading.

Sepal width and length seem to be relatively symmetrically distributed; petal length seems to be rather skewed, and petal width is somewhat skewed.

9 **Compare sepal length, sepal width, petal length, and petal width, using Figure 3.12.**

Iris Setosa: sepal length > sepal width > petal length > petal width

Versicolour: sepal length>petal length>sepal width>petal width

Virginica: sepal length>petal length>sepal width>petal width

17 **Discuss the differences between dimensionality reduction based on aggregation and dimensionality reduction based on techniques such as PCA and SVD.**

Aggregation, Principle Component Analysis, Singular Value Decomposition are the dimensionality reduction techniques.
In aggregation, group of dimensions are combined. The aggregation can be viewed as change of scale.
PCA or SVD techniques can be viewed as data projection onto a reduced set of dimensions. PCA is the most common application of SVD technique.