Data Science – Assignment 01

Name: **Sanio Luke Sebastian**

Roll No: **35**

Subject: **Data Science**

Date: **08-08-2022**

Topic: **Data Visualization**

Visualizing data is one of the most important techniques of data discovery and exploration. The discipline of data visualization encompasses the methods of expressing data in an abstract visual form. The visual representation of data provides easy comprehension of complex data with multiple attributes and their underlying relationships.

The motivation for using data visualization includes:

1. **Comprehension of dense information:** A simple visual chart can easily include thousands of data points. By using visuals, the user can see the big picture, as well as longer term trends that are extremely difficult to interpret purely by expressing data in numbers.

2. **Relationships**: Visualizing data in Cartesian coordinates enables exploration of the relationships between the attributes. Although representing more than three attributes on the x, y, and z-axes is not feasible in Cartesian coordinates, there are a few creative solutions available by changing properties like the size, color, and shape of data markers or using flow maps, where more than two attributes are used in a two-dimensional medium. Vision is one of the most powerful senses in the human body.

A. Univariate Visualization

Visual exploration starts with investigating one attribute at a time using univariate charts. The techniques like how the attribute values are distributed and the shape of the distribution.

1. ***Histogram***

   A histogram is one of the most basic visualization techniques to understand the frequency of the occurrence of values. It shows the distribution of the data by plotting the frequency of occurrence in a range. In a histogram, the attribute under inquiry is shown on the horizontal axis and the frequency of occurrence is on the vertical axis. For a continuous numeric data type, the range or binning value to group a range of values need to be specified.

   For example, in the case of human height in centimeters, all the occurrences between 152.00 and 152.99 are grouped under 152. Histograms are used to find the central location, range, and shape of distribution.

2. **Quartile**

   A box whisker plot is a **simple visual way of showing the distribution of a continuous variable** with information such as **quartiles, median, and outliers, overlaid by mean and standard deviation**. The main attraction of box whisker or quartile charts is that distributions of multiple attributes can be compared side by side and the overlap between them can be deduced.

3. **Distribution Chart**

   For continuous numeric attributes like petal length, instead of visualizing the actual data in the sample, its normal distribution function can be visualized instead. Here an inherent assumption is being made that the measurements of petal length (or any continuous variable) follow the normal distribution, and hence, its distribution can be visualized instead of the actual values. The normal distribution is also called the Gaussian distribution or "bell curve" due to its bell shape.

B. Multivariate Visualization

The **multivariate visual exploration considers more than one attribute in the same visual**. The techniques discussed in this section focus on the relationship of one attribute with another attribute. These visualizations examine two to four attributes simultaneously.

1. **Scatterplot**

   A **scatterplot** is one of the most powerful yet simple visual plots available. In a scatterplot, the data points are marked in Cartesian space with attributes of the dataset aligned with the coordinates. The attributes are usually of continuous data type. One of the key observations that can be concluded from a scatterplot is the existence of a relationship between two attributes under inquiry.

   If the attributes are linearly correlated, then the data points align closer to an imaginary straight line; if they are not correlated, the data points are scattered. Apart from basic correlation, scatterplots can also indicate the existence of patterns or groups of clusters in the data and identify outliers in the data. This is particularly useful for low-dimensional datasets.

2. **Scatter Multiple**
   A scatter multiple is an enhanced form of a simple scatterplot where more than two dimensions can be included in the chart and studied simultaneously. The primary attribute is used for the x-axis coordinate. The secondary axis is shared with more attributes or dimensions.

3. **Scatter Matrix**
   If the dataset has more than two attributes, it is important to look at combinations of all the attributes through a scatterplot. A scatter matrix solves this need by comparing all combinations of attributes with individual scatterplots and arranging these plots in a matrix.

   In effect, there are six distinct comparisons in scatter multiples of four attributes. Scatter matrices provide an effective visualization of comparative, multivariate, and high-density data.

4. **Bubble Chart**
   A bubble chart is a variation of a simple scatterplot with the addition of one more attribute, which is used to determine the size of the data point

5. **Density Chart**
   Density charts are similar to the scatterplots, with one more dimension included as a background color. The data point can also be colored to visualize one dimension, and hence, a total of four dimensions can be visualized in a density chart.

C. <u>Visualizing High-Dimensional Data</u>

Visualizing more than three attributes on a two-dimensional medium (like a paper or screen) is challenging. This limitation can be overcome by using transformation techniques to project the high-dimensional data points into parallel axis space. In this approach, a Cartesian axis is shared by more than one attribute.

1. **Parallel Chart**
   A parallel chart visualizes a data point quite innovatively by transforming or projecting multi-dimensional data into a two-dimensional chart medium. In this chart, every attribute or dimension is linearly arranged in one coordinate (x-axis) and all the measures are arranged in the other coordinate (y-axis). Since the x-axis is multivariate, each data point is represented as a line in a parallel space.

2. **Deviation Chart**
   deviation chart is very similar to a parallel chart, as it has parallel axes for all the attributes on the x-axis. Data points are extended across the dimensions as lines and there is one common y-axis. Instead of plotting all data lines, deviation charts only show the mean and standard deviation statistics.

3. **Andrews Curves**
   An Andrews plot belongs to a family of visualization techniques where the high-dimensional data are projected into a vector space so that each data point takes the form of a line or curve.