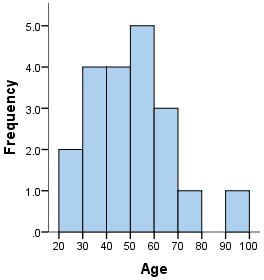
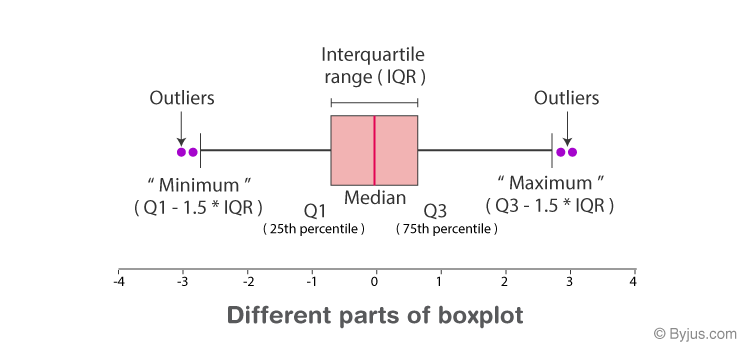
You may have heard the phrase, “Data is the new oil”. Like oil, data is very valuable but its value can't be used if we can not refine it. Refining oil gives us petrol, plastic, and beneficial products. Such is the case for data. Important information can be retrieved from a pile of data. The information then can be used by organizations to make a profit or by researchers to find groundbreaking discoveries. With the growing use of the internet, computing, and communication, the pile of data is larger than ever. There is no topic for which we can not get the data for. We can get the data for which medicine is used in certain pain and which movie to watch next. But data itself is not enough. Presenting raw data full of numbers and symbols to your group and trying to convince everyone in the room to jump on your idea because the data supports your concept can be troublesome. There must be some visualization of the information from the data. There are different ways and steps to get the desired information from a table of data. Data visualization is one of the most powerful tools for working with data.

Before understanding the visualization of data, we need to understand the type of data that can be present in the dataset. The kind of data plot we use depends on the data that we work on. Data can be of two types, continuous and categorical data. Continuous data are those data that can be represented by numbers. weight and height of a person are examples of continuous data. A person can weigh 50kg, 51kg, 51.5 kg, and so on. Categorical data are the data that can be classified. They are usually represented by text, but sometimes can be numbers too. A car can be categorized by the color Red, Blue, or Black. There are also other examples of categorical data with a longer list of categories. Finally, something can be of both categorical and continuous type. For example, age is a number such as 8,17,24 but it can be categorized by age group such as 10-15, 15-20, 20-25, and so on. We can treat this type of data either as continuous or categorical. It depends on the situation and operation. After understating the data we are working on, we can choose a proper plot to get insight into the dataset.

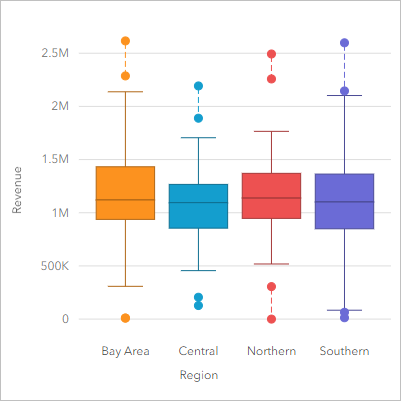
The first plot we are looking into is Histograms. Histograms are the type of plot that takes one continuous variable as input. A histogram plots numeric values as the bar of a series, representing the frequency of the range of the value in the data set. Each bar represents a range of numeric values called bin and the height of the gives the frequency of the data point within the range of bin. Histogram allows you to answer the question about the shape of the distribution of the variable. You can get the highest, lowest, and most common values in a data set from the histogram.



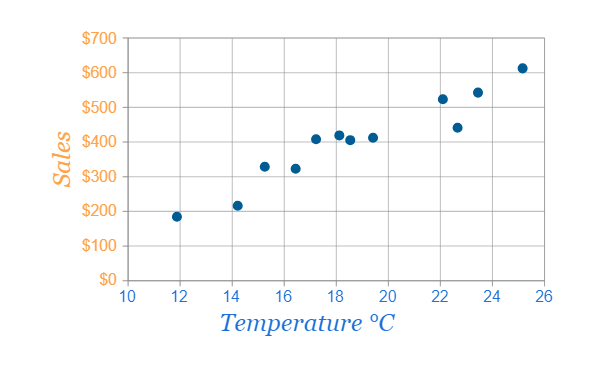
The figure shown is a histogram that visualizes the shape of the distribution of the Age variable. Age is represented by the x-axis. Bins of 10 years of width are used. On the y-axis we have frequency. Now, from the histogram, we can say that there are two records with ages in the range of 20-30 years and 1 record with age from 90-100. The most common range of age is 50-60 with frequency of 5. Individual histograms are great, but it is not efficient to compare many distributions at once. In this case, the box plot becomes more handy.

A box plot is a plot in which distributions of multiple categories of numeric data are represented using boxes, lines, and sometimes whiskers. Box plots are used to compare the distributions of numeric data values between multiple groups. They can be used to show the data distribution of a single group but it is specially used when you have to compare the data distribution among different groups. 

The box plot is based on a dataset’s quartiles. The ends of the box in the box plot mark two quartiles where the lower end marks the first quartile(Q1) and the third quartile (Q3) is marked by the box's upper end. The central line of the box mark the median or second quartile(Q2) of the dataset. The distance between Q3 and Q1 is known as the interquartile range (IQR). The whiskers extending from the box are within 1.5 times the IQR. Any data point further than the whisker is considered an outlier. From the box plot, we can get general information about the data’s symmetry, skew, variance, and outliers.

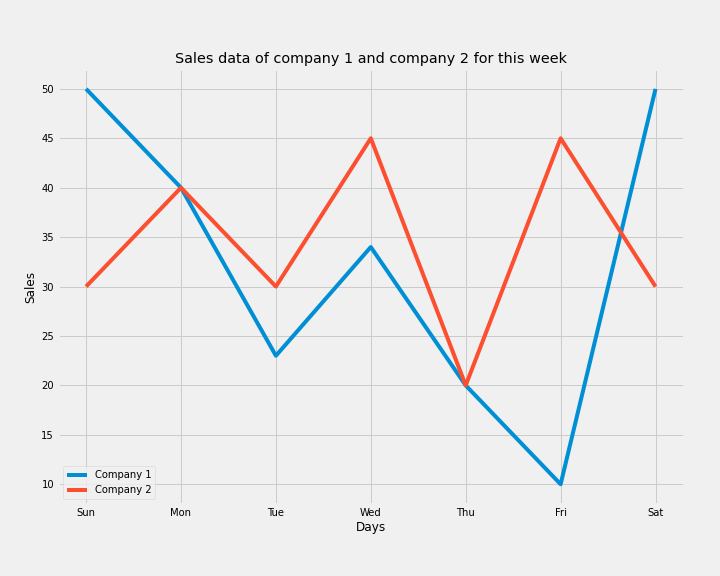


In the box plot, we can see the distribution of the numeric data revenue grouped together by region. From the plot, we can see revenue mostly lying around the value of 1M for all the regions. Southern and Bay Area regions has more distributed revenue data than of Central and Northern region. We can also see there are outliers among the grouped data with three of them being 0 revenue. There are also 2 outliers revenue with value being more than 2.5 M.

Up to now, we have been looking at the plot that shows the distribution of data of a variable numeric value. Now we will look into two variables. A scatter plot uses a dot to represent data points for two numeric values. Each of the variables are represented by vertical and horizontal lines. The position of the dot on the vertical and horizontal lines represents the values of the data point. A scatter plot should be used when you have two continuous variables, and you want to know about their relationship. For example, if one variable increases, does the other one increase too or does it decrease? The dot on the scatter plot not only gives the values of individual data points but also reports the pattern when the data is taken as a whole. One of the important concepts while interpreting scatter plots is correlation which expresses the extent to which two variables are linearly related.

In the given figure of the scatter plot, each dots show the sales and temperature at which the sale is made. Each dot represents a single interaction; each point’s horizontal position indicates the temperature at which the sale is made and the vertical position indicates the amount earned on the sale. From the plot, we can see a slight positive correlation between sales and temperature.

A close relative to a scatter plot is a line plot. Similar to a scatter plot, each data point is represented by vertical and horizontal lines. The position of the dot on the vertical and horizontal lines represents the values of the data point. These data points are connected by line segments to demonstrate the trend of change in value. Here, the horizontal line is often an independent variable. In most cases variable of Time is represented by a horizontal line while the vertical line reports the values for the required time. We should use a line plot when we want to report the changes in the value of a variable for continuous values of a second variable such as time. One of the useful features of a line plot is that we can draw multiple lines representing different subgroups in a single graph to compare them.



The given graph of a line plot demonstrates the number of sales made by two companies, company 1 represented by the blue line and company 2 represented by the red line across the 7 days of a week. Looking at the trend line, we can analyze that the sales company 1 fructuate from maximum to minimum sales. Meanwhile, company 2 demonstrated consistent sales across the 7 days. Looking at single data points for a day, company 1 made 50 sales on Sunday while company 2 made 30 sales. On Friday, company 1 only made sales of 10 while company 2 made sales of 45.

We cover some basics of data visualization with some commonly used plots along with their example. There is more to the understanding of data visualization. We barely scratch the surface of data visualization in this paper. There are more features that we can add to make the plot more informative and clean. For example, in a scatter plot, we can use colors or shapes for data points to report the third variable of the data set. We can also add a trend line to present the information clearly. For the histogram, there is to learn about skewness, kurtosis, size of the bins, and other features of the plot. Similarly, there are more things to learn about the box plots. Each type of plot has its features and parameters. Understanding these features can help us plot any given data with the best plot. We can further enhance our knowledge and skill in data visualization by learning more plots such as bar plots, heat maps, pie charts, and so on. We can learn data visualization using tools like Tableau, Chart.js, and Google Charts. We can also use Python with its vast packages and libraries. When we keep learning the data visualization, data will look easy.