Chapter 2
Visualization of Data

Objectives

After completing this chapter, you will be able to:

- To Create a Visualization of Data
- Understand the different types of data visualization
- Get the required information on data to be used on visualization
- Summary statistics data into visualize form

Introduction

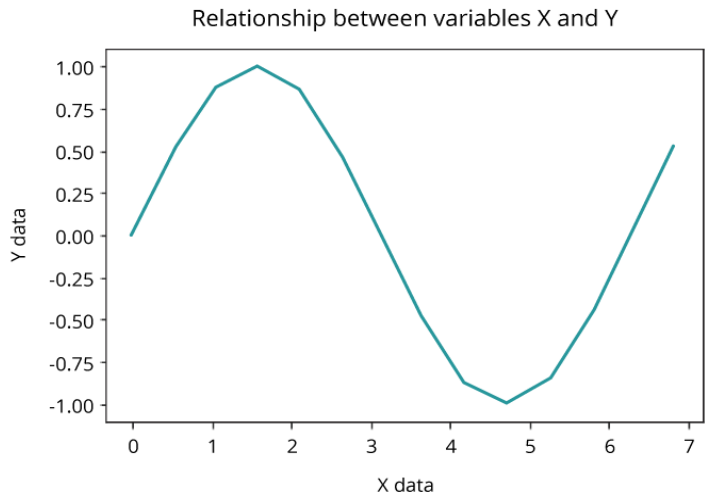
Researchers agree that vision is our dominant sense: 80–85% of information we perceive, learn or process is mediated through vision. It is even more so when we are trying to understand and interpret data or when we are looking for relationships among hundreds or thousands of variables to determine their relative importance. One of the most effective ways to discern important relationships is through advanced analysis and easy-to-understand visualizations.

Data visualization is applied in practically every field of knowledge. Scientists in various disciplines use computer techniques to model complex events and visualize phenomena that cannot be observed directly, such as weather patterns, medical conditions or mathematical relationships.

Data visualization provides an important suite of tools and techniques for gaining a qualitative understanding. The basic techniques are the following plots:

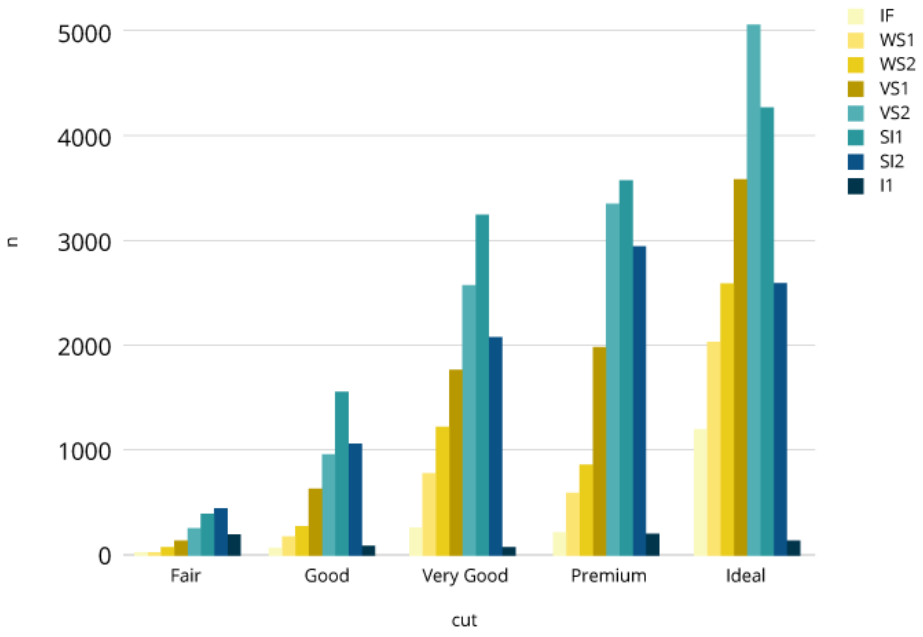
Line Plot

The simplest technique, a line plot is used to plot the relationship or dependence of one variable on another. To plot the relationship between the two variables, we can simply call the plot function.



Bar Chart

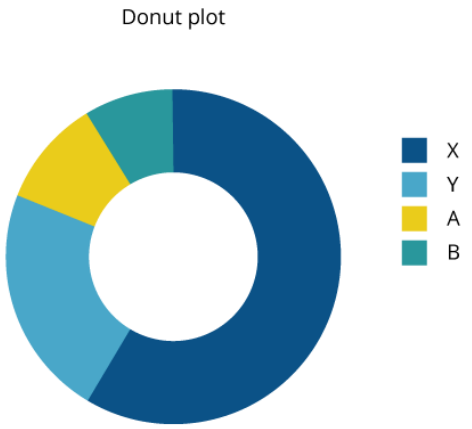
Bar charts are used for comparing the quantities of different categories or groups. Values of a category are represented with the help of bars and they can be configured with vertical or horizontal bars, with the length or height of each bar representing the value.



Pie and Donut Charts

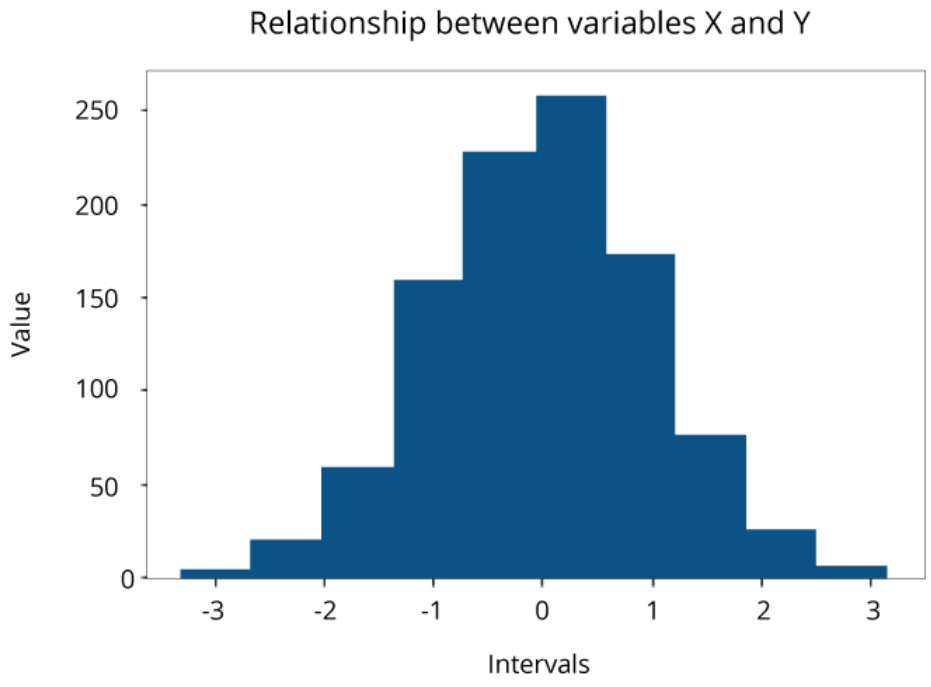
There is much debate around the value of pie and donut charts. As a rule, they are used to compare the parts of a whole and are most effective when there are limited components and when text and percentages are included to describe the content.

However, they can be difficult to interpret because the human eye has a hard time estimating areas and comparing visual angles.



Histogram Plot

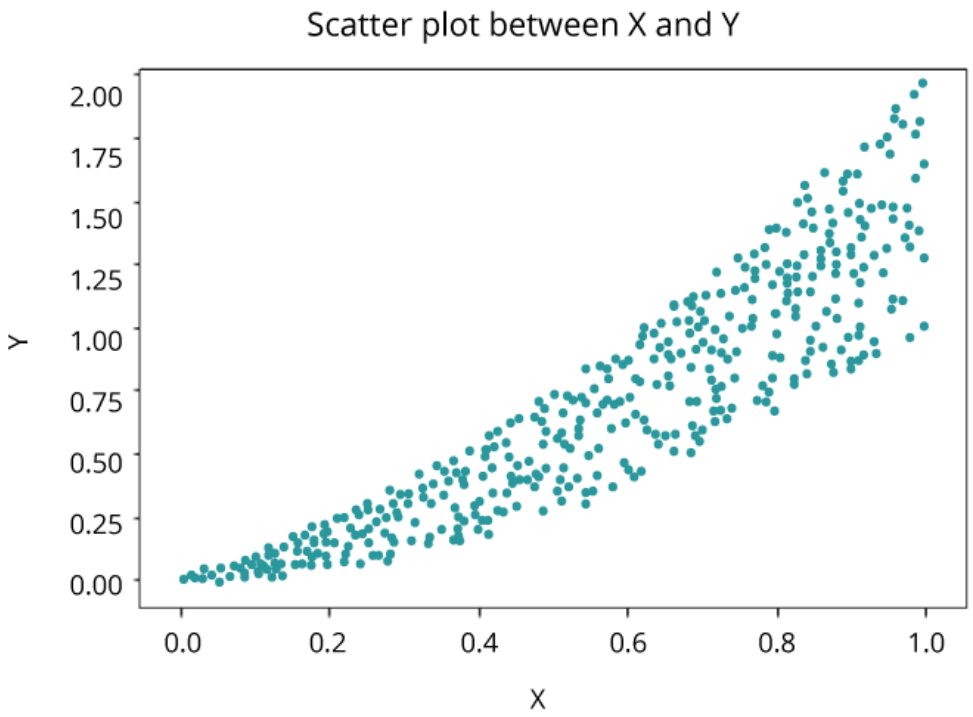
A histogram, representing the distribution of a continuous variable over a given interval or period of time, is one of the most frequently used data visualization techniques in machine learning. It plots the data by chunking it into intervals called 'bins'. It is used to inspect the underlying frequency distribution, outliers, skewness, and so on.



Scatter Plot

Another common visualization techniques is a scatter plot that is a two-dimensional plot representing the joint variation of two data items. Each marker

(symbols such as dots, squares and plus signs) represents an observation. The marker position indicates the value for each observation. When you assign more than two measures, a scatter plot matrix is produced that is a series of scatter plots displaying every possible pairing of the measures that are assigned to the visualization. Scatter plots are used for examining the relationship, or correlations, between X and Y variables.

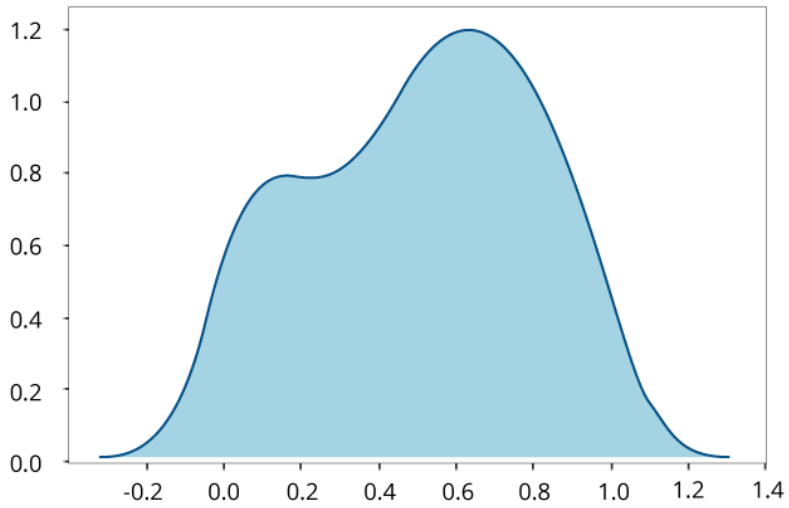


Visualizing Big Data

Today, organizations generate and collect data each minute. The huge amount of generated data, known as Big Data, brings new challenges to visualization because of the speed, size and diversity of information that must be taken into account. The volume, variety and velocity of such data requires from an organization to leave its comfort zone technologically to derive intelligence for effective decisions. New and more sophisticated visualization techniques based on core fundamentals of data analysis take into account not only the cardinality, but also the structure and the origin of such data.

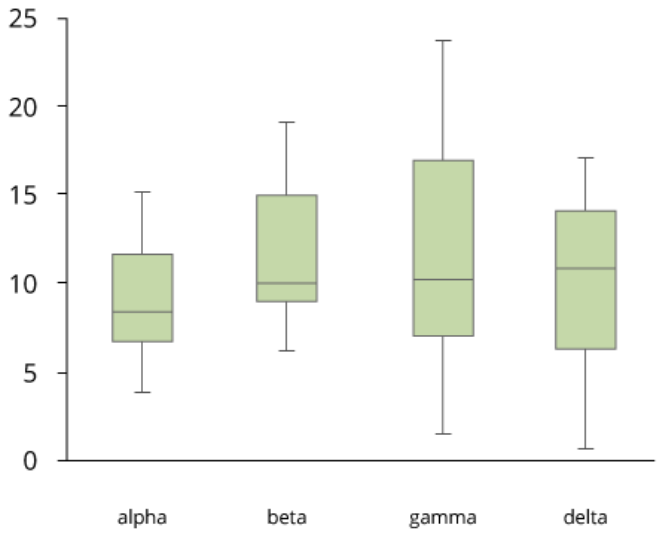
Kernel Density Estimation for Non-Parametric Data

If we have no knowledge about the population and the underlying distribution of data, such data is called non-parametric and is best visualized with the help of Kernel Density Function that represents the probability distribution function of a random variable. It is used when the parametric distribution of the data doesn't make much sense, and you want to avoid making assumptions about the data.



Box and Whisker Plot for Large Data

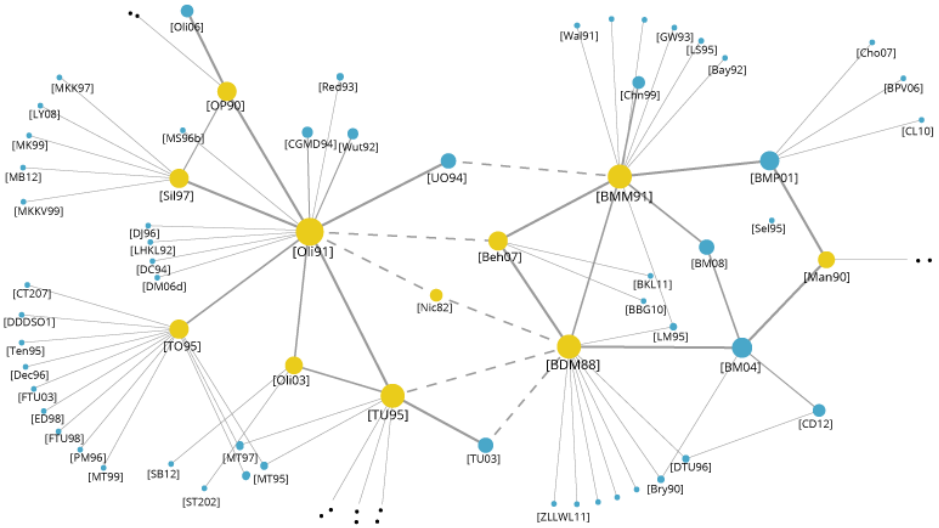
A binned box plot with whiskers shows the distribution of large data and easily sees outliers. In its essence, it is a graphical display of five statistics (the minimum, lower quartile, median, upper quartile and maximum) that summarizes the distribution of a set of data. The lower quartile (25th percentile) is represented by the lower edge of the box, and the upper quartile (75th percentile) is represented by the upper edge of the box. The median (50th percentile) is represented by a central line that divides the box into sections. Extreme values are represented by whiskers that extend out from the edges of the box. Box plots are often used to understand the outliers in the data.



Word Clouds and Network Diagrams for Unstructured Data

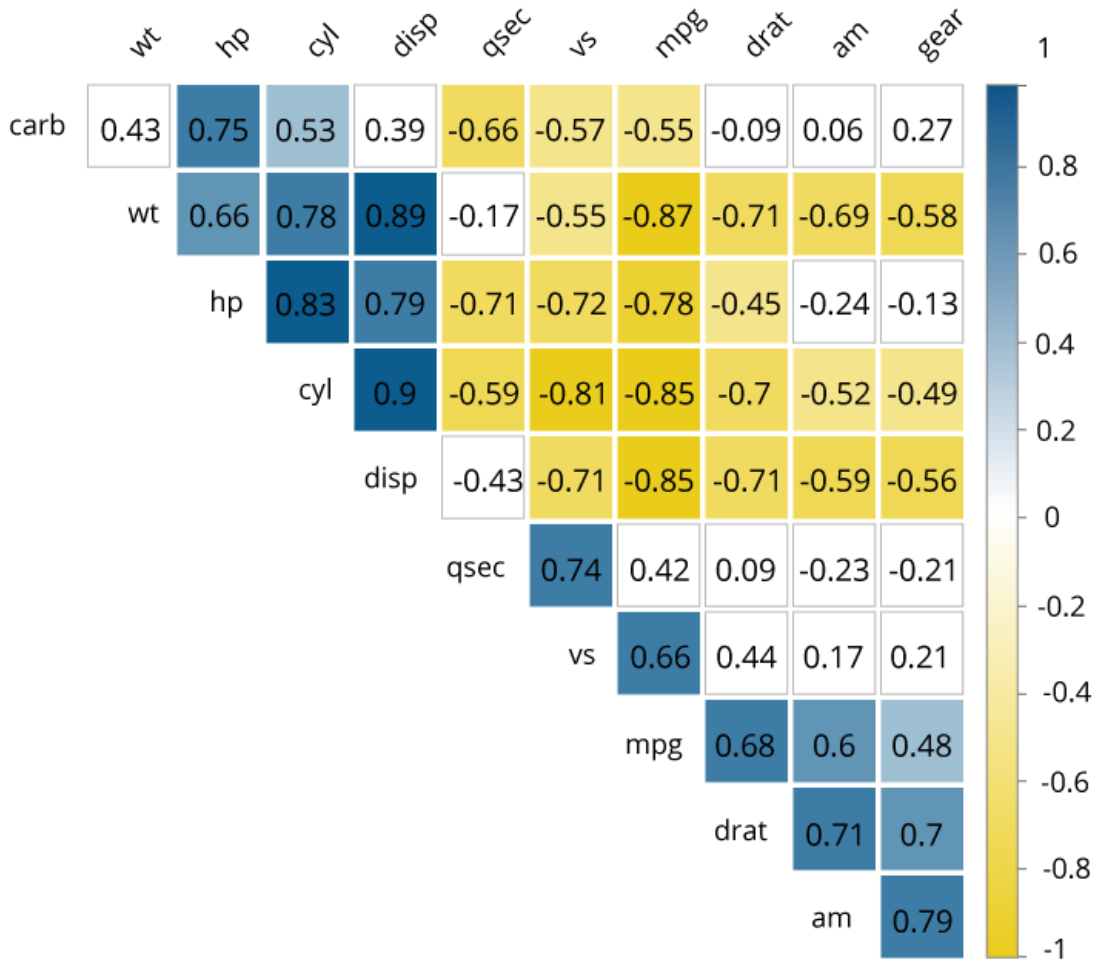
The variety of big data brings challenges because semistructured and unstructured data require new visualization techniques. A word cloud visual represents the frequency of a word within a body of text with its relative size in the cloud. This technique is used on unstructured data as a way to display high- or low-frequency words.

Another visualization technique that can be used for semistructured or unstructured data is the network diagram. Network diagrams represent relationships as nodes (individual actors within the network) and ties (relationships between the individuals). They are used in many applications, for example for analysis of social networks or mapping product sales across geographic areas.



Correlation Matrices

A correlation matrix allows quick identification of relationships between variables by combining big data and fast response times. Basically, a correlation matrix is a table showing correlation coefficients between variables: Each cell in the table represents the relationship between two variables. Correlation matrices are used as a way to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.



Data visualization may become a valuable addition to any presentation and the quickest path to understanding your data. Besides, the process of visualizing data can be both enjoyable and challenging. However, with the many techniques available, it is easy to end up presenting the information using a wrong tool. To choose the most appropriate visualization technique you need to understand the data, its type and composition, what information you are trying to convey to your audience, and how viewers process visual information. Sometimes, a simple line plot can do the task saving time and effort spent on trying to plot the data using advanced Big Data techniques. Understand your data—and it will open its hidden values to you.

A company which neglects proper descriptive statistics and data analysis already finds itself at a disadvantage compared to its competition. With data quickly turning into the lifeblood of the business world, it must be put to good use for a business to remain relevant and successful. The first step is collecting the data, but in many ways, that's the easy part. Once all that information has been gathered, what should companies do next? How does one make use of the large sets of data now at hand? That's where descriptive statistics and data visualization come in.

What is Descriptive Statistics?

The basic definition of descriptive statistics is that it describes data. It's a way to summarize and organize all that data you've collected into something more manageable and easy to understand. These descriptions can include either the entire dataset or merely a portion of it. One of the important things to know about descriptive data analysis is that it only focuses on the data itself and not on possible far-reaching implications that go well beyond the data that's represented. That's the difference in descriptive vs. inferential statistics, the latter of which uses more complex calculations to make wide-ranging predictions.

Descriptive Statistics Examples

To better understand the role of descriptive analysis, it's helpful to know some examples of descriptive statistics, and that starts with the types of statistical analysis you could encounter. The first type is the central tendency, mostly represented by the mean, median, or mode. The mean, of course, is the average of the dataset. The median is the value of the data point in the middle of the set. And the mode is the value which occurs most frequently. One common example of descriptive statistics related to mean is the student's grade point average (GPA). In this way, a student's academic performance can be measured by averaging his or her grades.

The second type is referred to as frequency. In other words, it's a measure of how frequently something happens. You've probably seen this in the descriptive statistics used in summarizing polls and survey responses — 61% of people answering "yes" to a specific question, for example.

The third type is a measure of position. This includes quartile and percentile ranks. Essentially, this type of descriptive statistical analysis helps to describe how different points of data relate to each other. The measure of position is best used to compare the data points to each other.

The fourth and final type of descriptive statistics is variation or dispersion. This type is most commonly used for determining the range of values that the data encompasses, identifying the maximum and minimum values in a descriptive statistics example. The variance of the information can also be attained through this approach, which can help determine if there are certain outliers among the data you've collected.

The Importance of Descriptive Statistics

Data analysis and descriptive statistics are vital components of any business strategy. Take a moment and think about what raw data looks like. It may come in the form of an enormous spreadsheet filled with numbers. Sometimes the data isn't even that organized. Even to data experts, this clutter of numbers can be hard to read much less interpret. Descriptive statistics organizes all of this information into something that is much more usable. This step usually has to be done before moving on to the next one — data visualization.

What is Data Visualization?

Like the term suggests, data visualizations is taking the data you have and converting it into a more visual form. Instead of having to look at numbers and spreadsheets, you get a picture that represents that information. While descriptive statistics can break data down into something more digestible, data visualization goes even further, taking that data and creating a visual that instantly communicates a story.

If you've ever seen a pie graph (and that's probably a given), then you know what this looks like in action. Pie graphs are very simple examples of visualization, but they're very effective in what they do. Think of bar charts, line graphs, spider charts, scatter plots, and diagrams and all the information they can convey in a moment. Think of it like the ultimate visual aid. It's easy to see why data visualization is a key ingredient in interpreting data.

The Importance of Data Visualization

From a business perspective, data visualization is indispensable. Data scientists may be able to look at raw data and discover key findings, but communicating what data says to those who lack expertise in data science will always be needed. If you need to get a point across in a short amount of time, data visualization is the way to do it. It makes the data clear and cohesive, eliminating the fluff and showing the most important points. With good data visualization, there will be no dispute over what the data is, rather the only discussion would be what to do with the data presented.

Data Visualization and Descriptive Statistics in Business

Combining both descriptive statistics and data visualization transforms them into a valuable asset for any company. One of the most important functions they serve is to help company leaders in making key business decisions. Data has normally been used when coming to a crucial business decision and the use of descriptive statistics and data visualization only amplifies that effectiveness.

There are many ways in which the two are used to inform business decisions. Through data visualization, it's easier to notice patterns and identify how various data points relate to each other. Business leaders can also look at recent historical trends and determine where those trends might go and how best the company can capitalize on them. With raw data, many of these instances would be hard to figure out, but after employing descriptive statistics and utilizing data visualization, the correlations can quickly become evident.

With these vital pieces, businesses suddenly become much more versatile. With the data visually displayed for everyone to understand, companies can identify untapped markets where their products or services might flourish. They can determine which parts of a company's operations can be made more efficient, thus cutting down on costs and optimizing overall performance. They may also identify ways to improve the customer experience by getting real time feedback from customers. Businesses can even prepare for future growth or possible downturns, keeping organizations ahead of the curve and ready to handle all the opportunities and challenges that await them.

The Right Data Visualization Tool

All of this requires the use of an effective and versatile data visualization tool, the exact kind that Import.io provides. With this tool, you can become proficient in understanding the data that you collect. A good data visualization tool like this is extremely helpful in turning abstract data into something much easier to grasp. As part of turning data into a visual element, the data gets cleaned, shaping into a manageable item and filtering out data values that may unnecessarily interfere with the message communicated through the information. Only through this process does data visualization turn data into something you can use to help strategize and plan ahead.

Use Data Today

Data analysis, descriptive statistics, and data visualization should become part of a business's arsenal. Data has so much to offer in terms of informing business decisions and planning business strategies. Missing out on these capabilities means missing out on possibilities and opportunities to grow and find greater success that's sustainable.

APPLYING VISUALIZATION TO STATISTICS ANALYSIS

One of the biggest challenges for visualization creation is the lack of a clear methodology for how to create a good visualization design. Graphic symbols are as vital a part of our communication systems as words in the language, numbers, and formulas in statistics. Many scholars discuss the subject of visualization and statistics. They report a similar problem—that visualization, unlike statistics, has no single methodology or single grammar syntax. You can visualize data in many different ways, from simple bar charts to more complex scatterplots or heat maps, based on your experience and your own ability. Adding to the confusion, there is no single comprehensive resource on the subject of data visualization, but rather an extensive array of books, journal articles, and online websites.

Stephanie Evergreen and Ann Emery

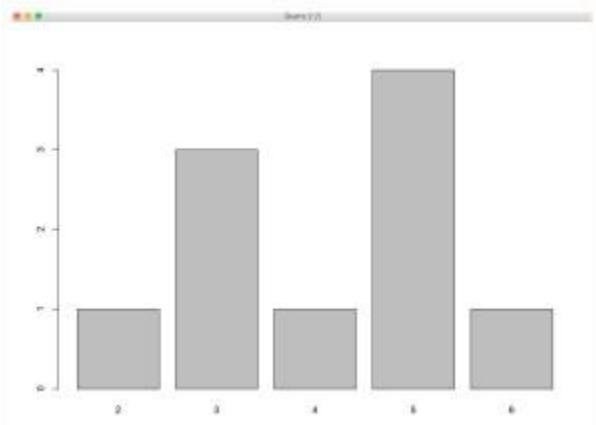
Stephanie Evergreen and Ann Emery (2014) provided a strategy checklist to enhance the user's experience for data visualization. The five key ideas when designing visualizations, according to Evergreen and Emery, consist of (1) supporting text description, (2) arrangement, (3) colors, (4) lines, and (5) overall meaning.

1. Supporting Text Description

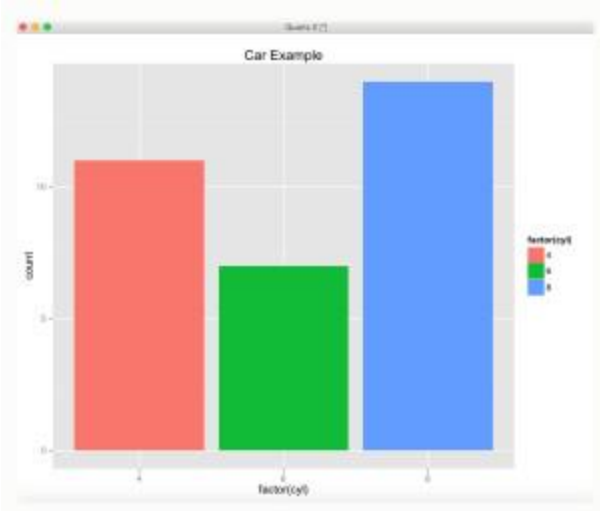
Adding a text description to support the visualization may help the user. The idea of adding a text description to the visualization is to clarify the graphics. According to Evergreen and Emery (2014),

>Use a six- to twelve-word descriptive title, left-justified in upper corner.
>Add subtitles and/or annotations to provide additional information.
>Lay out text horizontally. This includes titles, subtitles, annotations, and data labels. Line labels and axis labels can deviate from this rule and still receive full points.

The first example is without any captions.



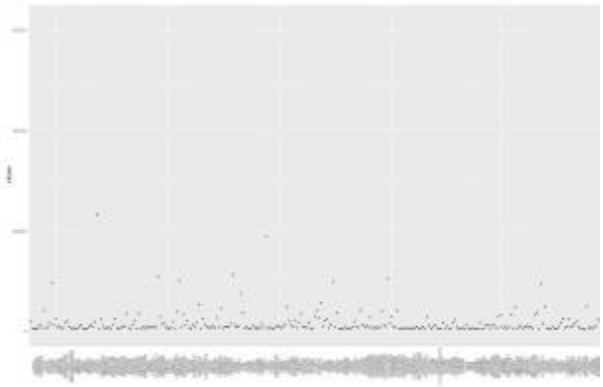
The second example is with captions.



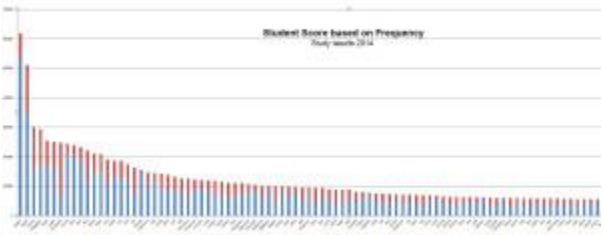
2. Arrangement

the next term Evergreen and Emery discuss is arrangement. Improper arrangement of graph elements can confuse readers at best and mislead viewers at worst. The goal of the arrangement is getting the viewer to focus on the substance of the visualization rather than on how the visualization was developed. We will illustrate the argument by providing two examples; the first example consists of disagreement where graph elements are not clearly outlined and the second example is with agreement.

i. Example of disagreement



ii. Example with agreement



3. Colors

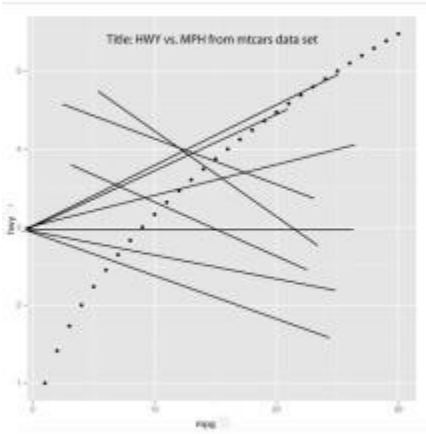
Colors are an important part of any visualization. We must think of colors when we

apply visualization to statistical analysis. Colors are the visual perceptual properties corresponding to the categories called red, blue, yellow, and others. Based on Evergreen and Emery (2014), colors are used to highlight key patterns. Action colors should guide the viewer to key parts of the display. Less important or supporting information should be in muted colors—mix your color arrangement with white or grey, making it less bright.

4. Lines

Lines are also an important part of the visualization. Excessive line use—gridlines, border tick marks, and axes can add clutter or noise to a graph, so eliminate them whenever they are not useful for interpreting data.

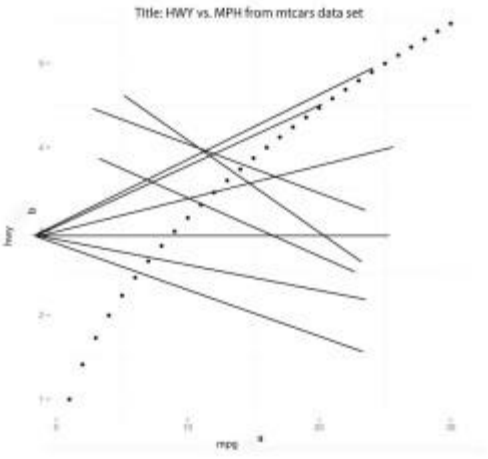
Our first example below consists of gridlines that, according to Evergreen and Emery, need to be muted.



In order to mute the background in ggplot2, we used the following code:

>theme(panel.border = element_blank())

The result:



Overall meaning:

While the meaning of visualization is still a difficult subject to determine, Evergreen and Emery recommend we provide more details in order to help the user to better understand the visualization. An important goal of any research scientist is the publication of the results of a completed study. Most academic and professional publications in our field require the researcher to provide a written document, based on their style that includes specifics of data analysis and data collection methods, in order for it to be accepted for review. Although the written word is still the dominant platform for reporting statistical

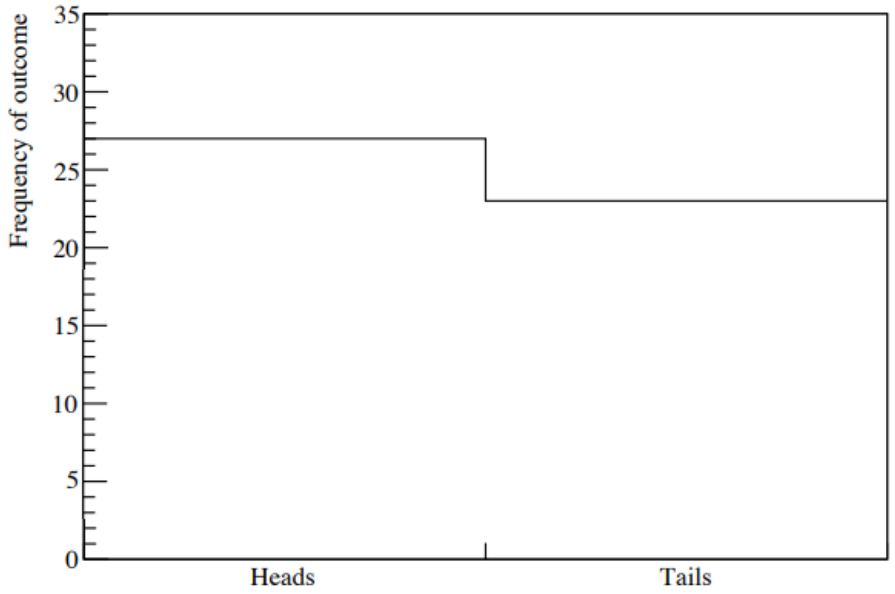
analysis results and recommendations, visualization has gained recognition. Researchers who employ statistical analyses in their studies often incorporate visualization graphics to help users see the results of the analysis.

Mode, Median, Mean

The **mean** is the average of the numbers. It is easy to calculate: add up all the numbers, then divide by how many numbers there are. In other words it is the sum divided by the count. The **mode** is the number that appears most frequently in a data set. A set of numbers may have one **mode**, more than one **mode**, or no **mode** at all. Other popular measures of central tendency include the mean, or the average (mean) of a set, and the median, the middle value in a set. **Median**: The middle number; found by ordering all data points and picking out the one in the middle (or if there are two middle numbers, taking the mean of those two numbers

Continuous data variables are discretized when represented by a histogram, in doing this one does loose information, so there has to be a trade off between the width of bins and the number of entries in a given bin. Indeed the width of all of the bins need not be constant when trying to find a balance between bin content and bin width.
If we take n repeated measurements of some observable x, it is useful to try and quantify our knowledge of the ensemble in terms of a single number to represent the value measured, and a second number to represent the spread of measurements. So our knowledge of the observable x will in general require some central value of the observable, some measure of the spread of the observable, and the units that the observable is measured in.

The mode of an ensemble of measurements is the most frequent value obtained. If the measurement in question is of a continuous variable, one has to bin the data in terms of a histogram in order to quantify the modal value of that distribution. The median value of the ensemble is the value of x where there are an equal number of measurements above and below that point. If there is an odd number of measurements, then this is straight forward and there are



The outcome of an ensemble of coin flipping experiments resulting in either Heads or Tails.

(n − 1)/2 points above and below the median value. If there is an even number of measurements, then the median value is taken as the midpoint between the two most central values. The median value can be useful when it is necessary to rank data (for example in computing upper limits or some correlation coefficients, described later in the course).

A better way to quantify the value measured is to take an arithmetic average of the individual measurements. The arithmetic mean value (usually this is just called the mean for short) of a set of data, is the average value of x computed from the ensemble The mean value is denoted either by x or (x) and is given by

$$\overline{x} = \langle x \rangle = \frac{1}{n} \sum_{i=1}^{n} x_i$$

where xi is the i th measurement of x. The mean value of a function f(x) can be calculated in the same way using

$$\overline{f} = \frac{1}{n} \sum_{i=1}^{n} f(x_i).$$

If the function in question is a continuous one, then the average value of the function between x = a and x = b is simply

$$\overline{f} = \frac{1}{b-a} \int_{x=a}^{b} f(x)dx.$$

It is possible to compute the average of a set of binned data, however if rounding occurs in the binning process, then some information is lost and the resulting average will be less precise than obtained using the above formulae

Video Links:

Visualization of Data

- https://www.youtube.com/watch?v=hEWY6kkBdpo
- https://www.youtube.com/watch?v=SKv7xUvJSpk

References

- https://datajournalism.com/read/handbook/one/understanding-data/using-data-visualization-to-find-insights-in-data
- https://www.kdnuggets.com/2019/04/best-data-visualization-techniques.html
- https://www.import.io/post/what-is-descriptive-statistics-and-how-data-visualization-can-transform-your-data/