# Chapter 5: Data Visualization

Data Visualization is the art and science of making data easy to understand and consume, for the end user. Ideal visualization shows the right amount of data, in the right order, in the right visual form, to convey the high priority information. The right visualization requires an understanding of the consumer's needs, nature of the data, and the many tools and techniques available to present data. The right visualization arises from a complete understanding of the totality of the situation. One should use visuals to tell a true, complete and fast-paced story.

Data visualization is the last step in the data life cycle. This is where the data is processed for presentation in an easy-to-consume manner to the right audience for the right purpose. The data should be converted into a language and format that is best preferred and understood by the consumer of data. The presentation should aim to highlight the insights from the data in an actionable manner. If the data is presented in too much detail, then the consumer of that data might lose interest and the insight.

**Caselet: Dr Hans Gosling - Visualizing Global Public Health**

*Dr. Hans Rosling is a master at data visualization. He has perfected the art of showing data in novel ways to highlight unexpected truths. He has become an online star by using data visualizations to make serious points about global health policy and development. Using novel ways to illustrate data obtained from UN agencies, he has helped demonstrate the progress that the world has made in improving public health on many dimensions. The best way to grasp the power of his work is to click here to see this TED video, where Life Expectancy is mapped along with Fertility Rate for all countries from 1962 to 2003. Figure 5.1 shows a one graphic from this video.*

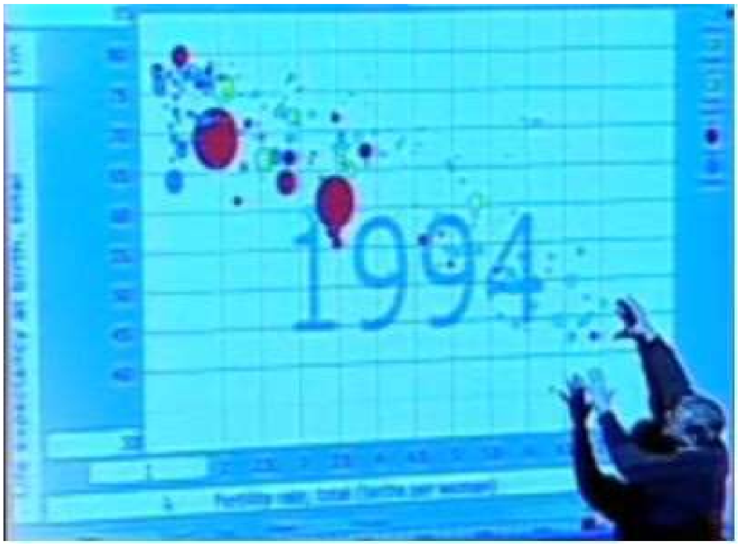

**Figure 5.1: Visualizing Global Health Data (source: ted.com)**

*"THE biggest myth is that if we save all the poor kids, we will destroy the planet," says Hans Rosling, a doctor and professor of international health at the Karolinska Institute in Sweden. "But you can't stop population growth by letting poor children die." He has the computerised graphs to prove it: colourful visuals with circles that swarm, swell and shrink like living creatures. Dr Rosling's mesmerizing graphics have been impressing audiences on the international lecture circuit, from the TED conferences to the World Economic Forum at Davos. Instead of bar charts and histograms, Dr Rosling uses*

*Lego bricks, IKEA boxes and data-visualization software developed by his Gapminder Foundation to transform reams of economic and public-health data into gripping stories. His aim is ambitious. "I produce a road map for the modern world," he says. "Where people want to drive is up to them. But I have the idea that if they have a proper road map and know what the global realities are, they'll make better decisions." (source: economist.com).*

*Q1: What are the business and social implications of this kind of data visualization?*
*Q2: How could these techniques be applied in your organization and area of work?*

## Excellence in Visualization

Data can be presented in the form of rectangular *tables*, or it can be presented in colorful graphs of various types. “Small, non-comparative, highly-labeled data sets usually belong in tables” – (Ed Tufte, 2001, p 33). However, as the amount of data grows, graphs are preferable. Graphics help give shape to data. Tufte, a pioneering expert on data visualization, presents the following objectives for graphical excellence:

1. *Show, and even reveal, the data*: The data should tell a story, especially a story hidden in large masses of data. However, reveal the data in context, so the story is correctly told.
2. *Induce the viewer to think of the substance of the data*: The format of the graph should be so natural to the data, that it hides itself and lets data shine.
3. *Avoid distorting what the data have to say*: Statistics can be used to lie. In the name of simplifying, some crucial context could be removed leading to distorted communication.
4. *Make large data sets coherent*: By giving shape to data, visualizations can help bring the data together to tell a comprehensive story.
5. *Encourage the eyes to compare different pieces of data*: Organize the chart in ways the eyes would naturally move to derive insights from the graph.
6. *Reveal the data at several levels of detail*: Graphs leads to insights, which raise further curiosity, and thus presentations should help get to the root cause.
7. *Serve a reasonably clear purpose* – informing or decision-making.
8. *Closely integrate with the statistical and verbal descriptions of the dataset*: There should be no separation of charts and text in presentation. Each mode should tell a complete story. Intersperse text with the map/graphic to highlight the main insights.

Context is important in interpreting graphics. Perception of the chart is as important as the actual charts. Do not ignore the intelligence or the biases of the reader. Keep the template consistent, and only show variations in data. There can be many excuses for graphical distortion. E.g. “we are just approximating.” Quality of information transmission comes prior to aesthetics of chart. Leaving out the contextual data can be misleading.

A lot of graphics are published because they serve a particular cause or a point of view. It is particularly important when in a for-profit or politically contested environments. Many related dimensions can be folded into a graph. The more the dimensions that are represented in a graph, the richer and more useful the chart become. The data visualizer should understand the client’s objects and present the data for accurate perception of the totality of the situation.

## Types of Charts

There are many kinds of data as seen in the caselet above. Time series data is the most popular form of data. It helps reveal patterns over time. However, data could be organized around alphabetical list of things, such as countries or products or salespeople. Figure 5.2 shows some of the popular chart types and their usage.

1. *Line graph*. This is a basic and most popular type of displaying information. It shows data as a series of points connected by straight line segments. If mining with time-series data, time is usually shown on the x-axis. Multiple variables can be represented on the same scale on y-axis to compare of the line graphs of all the variables.
2. *Scatter plot:* This is another very basic and useful graphic form. It helps reveal the relationship between two variables. In the above caselet, it shows two dimensions: Life Expectancy and Fertility Rate. Unlike in a line graph, there are no line segments connecting the points.
3. *Bar graph:* A bar graph shows thin colorful rectangular bars with their lengths being proportional to the values represented. The bars can be plotted vertically or horizontally. The bar graphs use a lot of more ink than the line graph and should be used when line graphs are inadequate.
4. *Stacked Bar graphs:* These are a particular method of doing bar graphs. Values of multiple variables are stacked one on top of the other to tell an interesting story. Bars can also be normalized such as the total height of every bar is equal, so it can show the relative composition of each bar.
5. *Histograms*: These are like bar graphs, except that they are useful in showing data frequencies or data values on classes (or ranges) of a numerical variable.

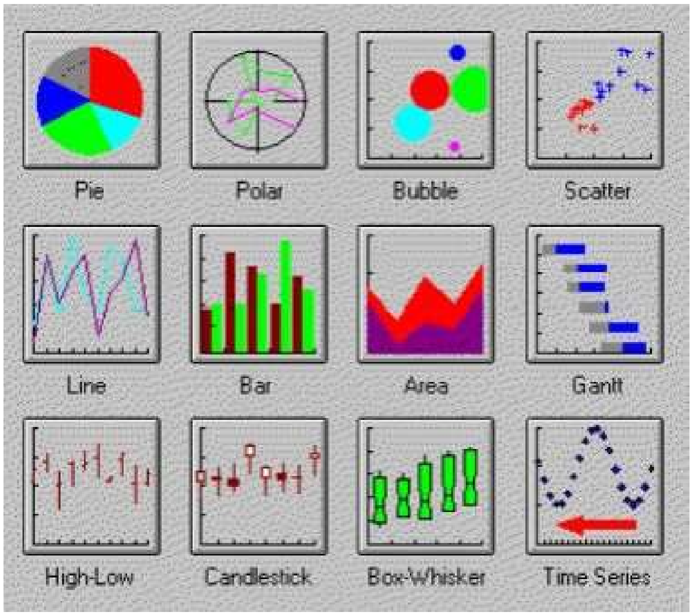

Figure 5.1: Many types of graphs

6. *Pie charts*: These are very popular to show the distribution of a variable, such as sales by region. The size of a slice is representative of the relative strengths of each value.
7. *Box charts:* These are special form of charts to show the distribution of variables. The box shows the middle half of the values, while whiskers on both sides extend to the extreme values in either direction.
8. *Bubble Graph*: This is an interesting way of displaying multiple dimensions in one chart. It is a variant of a scatter plot with many data points marked on two dimensions. Now imagine that each data point on the graph is a bubble (or a circle) … the size of the circle and the color fill in the circle could represent two additional dimensions.
9. *Dials*: These are charts like the speed dial in the car, that shows whether the variable value (such as sales number) is in the low range, medium range, or high range. These ranges could be colored red, yellow and gree to give an instant view of the data.
10. *Geographical* Data maps are particularly useful maps to denote statistics. Figure 5.3 shows a tweet density map of the US. It shows where the tweets emerge from in the US.

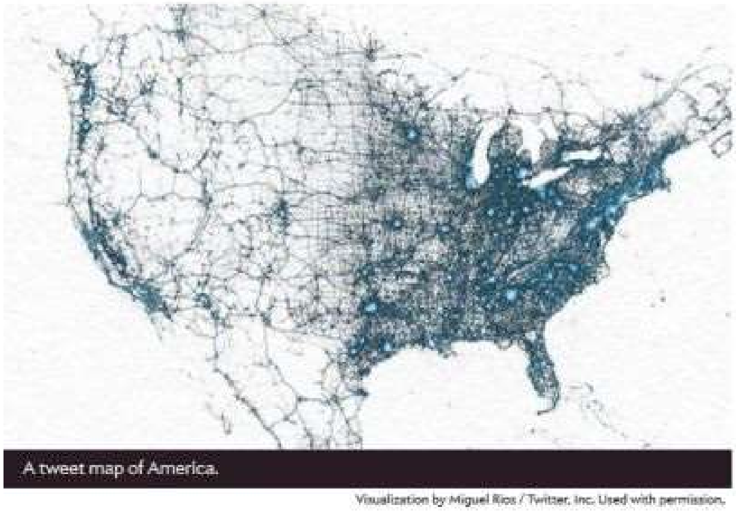
A tweet map of America.
Visualization by Miguel Rios / Twitter, Inc. Used with permission.

Figure 5.3: US tweet map (Source: Slate.com)

11. *Pictographs*: One can use pictures to represent data. E.g. Figure 5.2 shows the number of liters of water needed to produce one pound of each of the products, where images are used to show the product for easy reference. Each droplet of water also represents 50 liters of water.

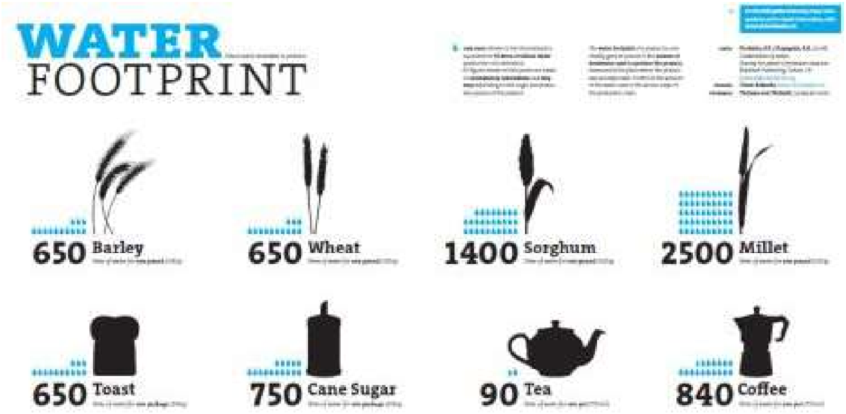

Figure 5.4: Pictograph of Water footprint (source : waterfootprint.org)

## Visualization Example

To demonstrate how each of the visualization tools could be used, imagine an executive for a company who wants to analyze the sales performance of his division. Figure 5.1 show the important raw sales data for the current year, alphabetically sorted by Product names.

| Product | Revenue | Orders | SalesPers |
|---|---|---|---|
| AA | 9731 | 131 | 23 |
| BB | 355 | 43 | 8 |
| CC | 992 | 32 | 6 |
| DD | 125 | 31 | 4 |
| EE | 933 | 30 | 7 |
| FF | 676 | 35 | 6 |
| GG | 1411 | 128 | 13 |
| HH | 5116 | 132 | 38 |
| JJ | 215 | 7 | 2 |
| KK | 3833 | 122 | 50 |
| LL | 1348 | 15 | 7 |
| MM | 1201 | 28 | 13 |

Table 5.1: Raw Performance Data

To reveal some meaningful pattern, a good first step would be to sort the table by Product revenue, with highest revenue first. We could total up the values of Revenue, Orders, and Sales persons for all products. We can also add some important ratios to the right of the table (Table 5.2).

| **Product** | **Revenue** | **Orders** | **SalesPers** | *Rev/Order* | *Rev/SalesP* | *Orders/SalesP* |
|---|---|---|---|---|---|---|
| AA | 9731 | 131 | 23 | *74.3* | *423.1* | *5.7* |
| HH | 5116 | 132 | 38 | *38.8* | *134.6* | *3.5* |
| KK | 3833 | 122 | 50 | *31.4* | *76.7* | *2.4* |
| GG | 1411 | 128 | 13 | *11.0* | *108.5* | *9.8* |
| LL | 1348 | 15 | 7 | *89.9* | *192.6* | *2.1* |
| MM | 1201 | 28 | 13 | *42.9* | *92.4* | *2.2* |
| CC | 992 | 32 | 6 | *31.0* | *165.3* | *5.3* |
| EE | 933 | 30 | 7 | *31.1* | *133.3* | *4.3* |
| FF | 676 | 35 | 6 | *19.3* | *112.7* | *5.8* |
| BB | 355 | 43 | 8 | *8.3* | *44.4* | *5.4* |
| JJ | 215 | 7 | 2 | *30.7* | *107.5* | *3.5* |
| DD | 125 | 31 | 4 | *4.0* | *31.3* | *7.8* |
| **Total** | **25936** | **734** | **177** | *35.3* | *146.5* | *4.1* |

**Table 5.2: Sorted data, with additional ratios**

There are too many numbers on this table to visualize any trends in them. The numbers are in different scales so plotting them on the same chart would not be easy. E.g. the Revenue numbers are in thousands while the SalesPers numbers and Orders/SalesPers are in the single or double digit.

One could start by visualizing the revenue as a pie-chart. The revenue proportion drops significantly from the first product to the next. (Figure 5.5). It is interesting to note that the top 3 products produce almost 75% of the revenue.

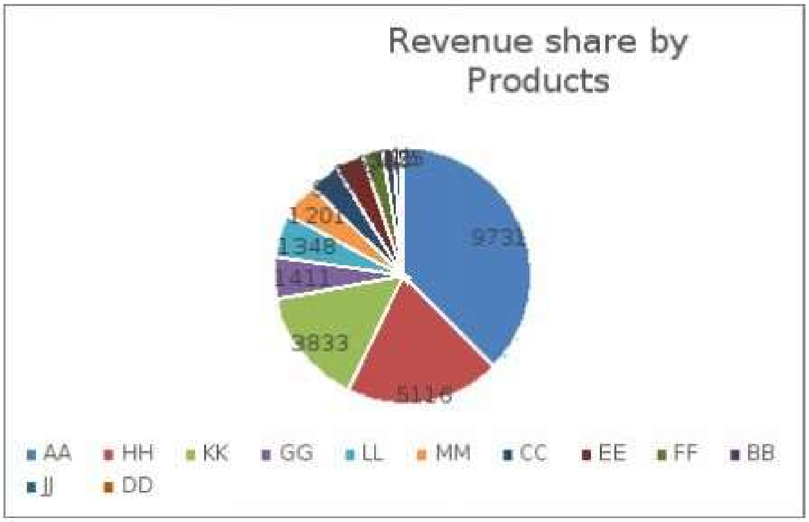

Figure 5.5: Revenue Share by Product

The number of orders for each product can be plotted as a bar graph (Figure 5.2). This shows that while the revenue is widely different for the top four products, they have approximately the same number of orders.

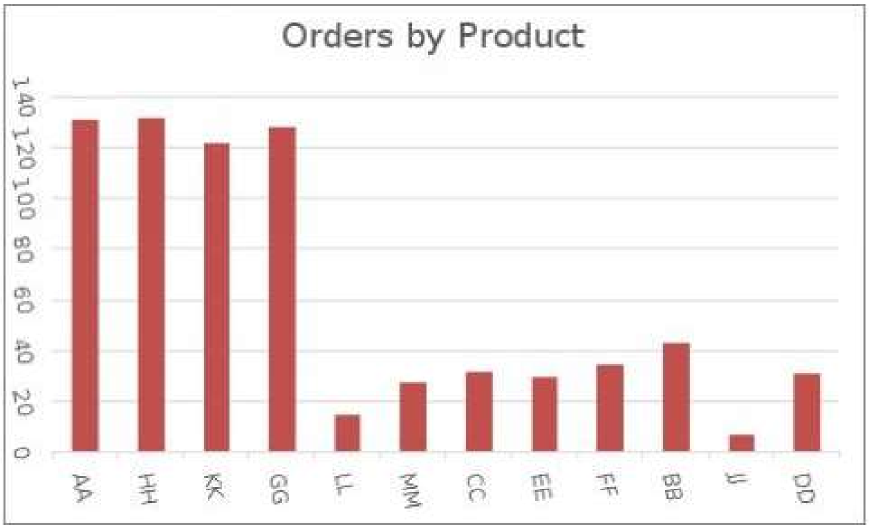

Figure 5.6: Orders by Products

Therefore, the orders data could be investigated further to see order patterns. Suppose additional data is made available for Orders by their size. Suppose the orders are chunked into 4 sizes: Tiny, Small, Medium, and Large. Additional data is shown in Table 5.3.

| Product | Total Orders | Tiny | Small | Medium | Large |
|---|---|---|---|---|---|
| AA | 131 | 5 | 44 | 70 | 12 |
| HH | 132 | 38 | 60 | 30 | 4 |
| KK | 122 | 20 | 50 | 44 | 8 |
| GG | 128 | 52 | 70 | 6 | 0 |
| LL | 15 | 2 | 3 | 5 | 5 |
| MM | 28 | 8 | 12 | 6 | 2 |
| CC | 32 | 5 | 17 | 10 | 0 |
| EE | 30 | 6 | 14 | 10 | 0 |
| FF | 35 | 10 | 22 | 3 | 0 |
| BB | 43 | 18 | 25 | 0 | 0 |
| JJ | 7 | 4 | 2 | 1 | 0 |
| DD | 31 | 21 | 10 | 0 | 0 |
| Total | 734 | 189 | 329 | 185 | 31 |

**Table 5.3: Additional data on order sizes**

Figure 5.7 is a stacked bar graph that shows the percentage of Orders by size for each product. This chart (Figure 5.7) brings a different set of insights. It shows that the product HH has a larger proportion of tiny orders. The products at the far right have a large number of tiny orders and very few large orders.

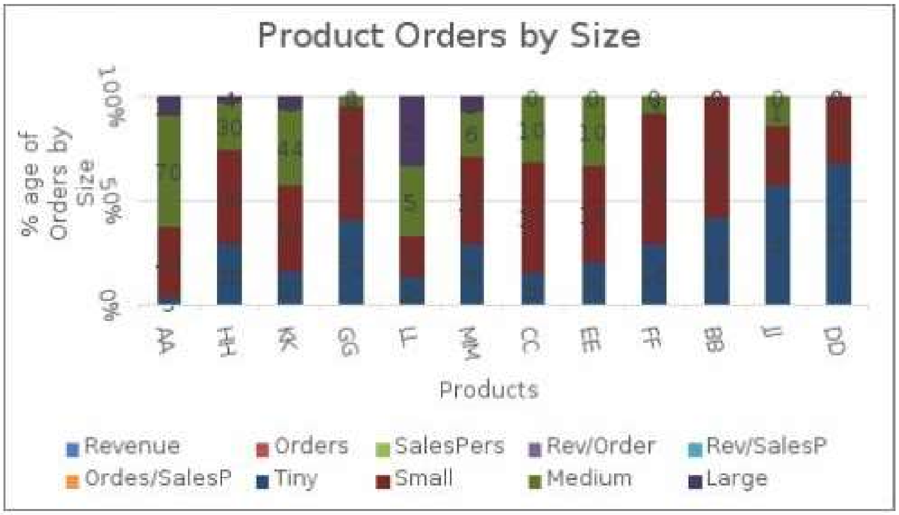

Figure 5.7: Product Orders by Order Size

## Visualization Example phase -2

The executive wants to understand the productivity of salespersons. This analysis could be done both in terms of the number of orders, or revenue, per salesperson. There could be two separate graphs, one for the number of orders per salesperson, and the other for the revenue per salesperson. However, an interesting way is to plot both measures on the same graph to give a more complete picture. This can be done even when the two data have different scales. The data is here resorted by number of orders per salesperson.

Figure 5.8 shows two line graphs superimposed upon each other. One line shows the revenue per salesperson, while the other shows the number of orders per salesperson. It shows that the highest productivity of 5.3 orders per sales person, down to 2.1 orders per salesperson. The second line, the blue line shows the revenue per sales person for each for the products. The revenue per salesperson is highest at 630, while it is lowest at just 30.

And thus additional layers of data visualization can go on for this data set.

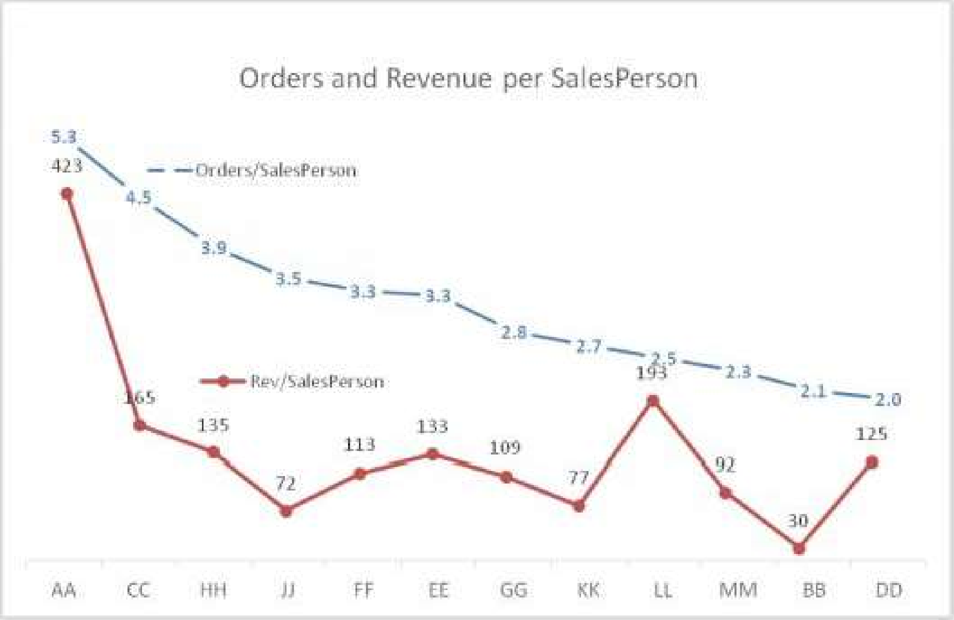

**Figure 5.8: Salesperson productivity by product**

## Tips for Data Visualization

To help the client in understanding the situation, the following considerations are important:

1. *Fetch appropriate and correct data for analysis*. This requires some understanding of the domain of the client and what is important for the client. E.g. in a business setting, one may need to understand the many measure of profitability and productivity.
2. *Sort the data in the most appropriate manner*. It could be sorted by numerical variables, or alphabetically by name.
3. *Choose appropriate method to present the data*. The data could be presented as a table, or it could be presented as any of the graph types.
4. *The data set could be pruned* to include only the more significant elements. More data is not necessarily better, unless it makes the most significant impact on the situation.
5. *The visualization could show additional dimension for reference* such as the expectations or targets with which to compare the results.
6. *The numerical data may need to be binned into a few categories*. E.g. the orders per person were plotted as actual values, while the order sizes were binned into 4 categorical choices.
7. *High-level visualization could be backed by more detailed analysis*. For the most significant results, a drill-down may be required.
8. *There may be need to present additional textual information* to tell the whole story. For example, one may require notes to explain some extraordinary results.

## Conclusion

Data Visualization is the last phase of the data lifecycle, and leads to the consumption of data by the end user. It should tell an accurate, complete and simple story backed by date, while keeping it insightful and engaging. There are innumerable types of visual graphing techniques available for visualizing data. The choice of the right tools requires a good understanding of the business domain, the data set and the client needs. There is ample room for creativity to design ever more compelling data visualization to most efficiently convey the insights from the data.

## Review Questions

1. What is data visualization?
2. How would you judge the quality of data visualizations?
3. What are the data visualization techniques? When would you use tables or graphs?
4. Describe some key steps in data visualization.
5. What are some key requirements for good visualization.

### Liberty Stores Case Exercise: Step 4

*Liberty is constantly evaluating its performance for improving efficiencies in all its operations, including the commercial operations as well its charitable activities.*

1. *What data visualization techniques would you use to help understand sales patterns?*
2. *What data visualization technique would you use to categorize its customers?*