Creating Business Value through Data Presentation

Introduction to Analyzing Data for Business Goals

# Presenting and Communicating Data

Business problems are often stated in the form of questions asked by stakeholders seeking solutions or needing to make business decisions. These stakeholders can be product managers, marketing directors, business analysts, or others involved in determining an organization’s strategy and actions, and as an analyst, your job is to provide them with information that allows them to implement effective solutions or make appropriate decisions. This requires accurate and relevant information, which comes in the form of data. Knowing the questions your stakeholders are asking helps you identify the specific data you’ll need to collect, analyze, and ultimately present.

Once you have acquired your data, you need to analyze it and present it to your stakeholders using suitable visualizations to help clarify the message you wish to communicate. These visualizations can include line graphs, bar charts, histograms, maps, and infographics produced by a business intelligence tool or other application, and can depict both objective (quantitative) and subjective (dependent on interpretation) values. As you craft your message, you must think about the effect you’re trying to achieve, whether the use of the data (or the message itself) is ethical, and who exactly is in your audience. Finally, depending on the business need you may want to consider the frequency with which you need to make presentations.

# Data Manipulation

When you first receive data from its original source (such as from a web form, hospital patient database, census tally, etc.), it is in raw form. Before you can use it to answer questions or solve business problems, you need to manipulate it—to modify it—so it can be used with analytical tools such as statistics applications or presented in either a visual or tabular form for direct consumption by a specific audience. These modifications involve trimming or truncating data elements, sorting the data according to specified criteria, merging data fields, transforming specified rows or columns, and so forth.

Data typically comes in a variety of forms, including numbers, text strings, tables, images, etc. Before you can use it, you must prepare it for input into the software applications you want to use to analyze it. This may require the data to be converted into Microsoft Excel format for use in spreadsheets or CSV (comma-separated variable) files for use in a statistics package. You also need to ensure that the data is error-free. There must be no missing values and no obviously “wrong” data elements.

For example, you might notice errors in the first row of a data table and decide to eliminate the entire row. You might see a column that contains data that isn’t important for your analysis, so you eliminate it. You might want to change negative numbers in a column to positive numbers because you plan to take the square root of those values in an analytical calculation. Perhaps dates appear in the format *year/month/date* in the raw data but your analysis software requires dates to be input as *date/month/year*, so you’ll need to manipulate the order of those date elements before exporting them to the analysis application. You may also want to delete sensitive data such as social security numbers before passing the data set on for further work. These are the kinds of manipulations you need to make so the data can be used by analytics or reporting packages.

## Input/Output: File Read and Write

There are many tools available for inputting enterprise data through *extraction, transformation,* and *loading* (ETL) and *extraction, loading*, and *transformation* (ELT). These two processes sound almost the same but are actually quite different. In ETL, the data is first transferred from its source to a staging area before being sent to its destination, which is usually a data warehouse. In ELT, the staging step is skipped and the transformation takes place within the data warehouse itself. One can also adopt a simple approach such as using Python or Excel where one line of Python code or the menu options in Excel can initiate the transfer. The data itself can be in one of a variety of formats including CSV or XML (extensible markup language). Finally, data can be ingested more directly via an application programming interface (API).

Upon completion, the results can be output back into the original file or into a new file. This last step is important because we’ll often have large amounts of data that can’t be processed in a computer’s working memory or that you wish to store for later use or sharing with other analysts.

## Data Inspection

To determine the changes that need to be made to raw data (if any), you must carefully inspect the data and develop a full understanding of it. You look at samples of rows and columns. You look at structures within the data. You look for metadata (what the data tells you about itself). You assess the condition of the data. Is there any missing data? (For instance, does a time series contain only six months of data where you need a full year’s worth?) Are there any relationships among the data in several columns? Are the relationships linear or nonlinear? For example, if the data in the two columns are linearly correlated, you may not need to keep both columns. How about groups or subgroups? Are any of these missing bigger “chunks” of data than others? Datasets have facts, dimensions, hierarchies, possible aliases, and various attributes. Although it’s possible to engage in data inspection at any time, it’s usually better to do it when you first begin working with the data.

### Missing Values Data Operations

When dealing with missing values, always go back to the overall goal of your analysis. Are you looking to create predictive models? Visualizations? What questions are you trying to answer? With a clear vision of your objectives, you can see which types of errors and missing data either won’t affect your analysis or can be “filled in” by other means.

Your intention, then, is to replace missing values with reasonable estimates. One way of doing this is by interpolation using adjacent values. Another way is to use an average of historical values. You can also apply more sophisticated approaches such as linear regression or probabilistic modeling to estimate the missing values. As it turns out, there are quite a few ways to find appropriate estimates.

Regarding errors, it may not always be possible to completely eliminate them so you could run your analysis both with and without them to determine their impact. Ideally, it’s best to catch errors as early as possible in the analysis process. Of course, if there are too many missing or erroneous data values, the dataset may not be usable. Making a decision about a dataset’s usability is purely subjective based on the particular situation at hand.

## Tools and Techniques

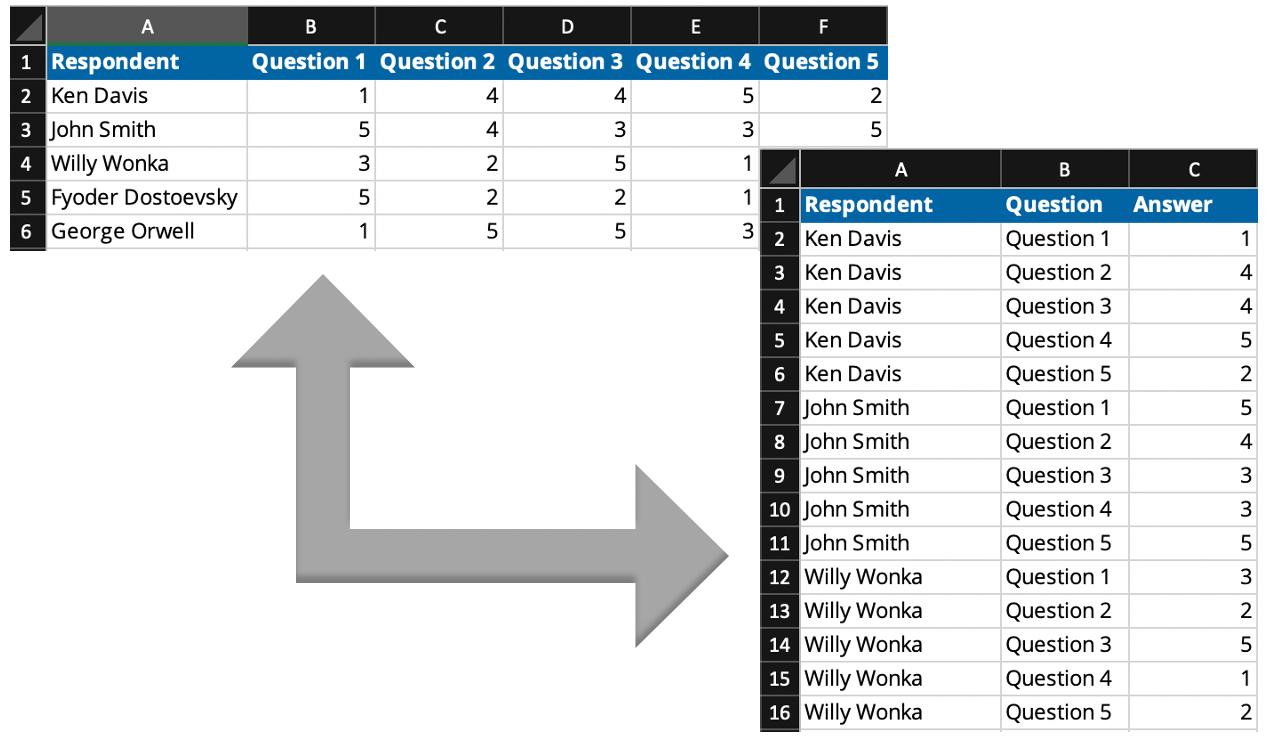
Data often needs to be moved, rotated, or transformed in some form or another. Why? There are many possible reasons; here are three:

* The software tool you’re using to analyze the data or organize it for a presentation may accept the values for input only in a particular sequence or format. For instance, you may need to switch items from columns to rows, or vice-versa.
* You may want to insert new data and thus need to shift the existing data. For example, if you insert a new data field into your table, you’ll need to shift existing columns to the right.
* You may want to re-sort the data so specific rows appear higher in the table. For instance, if you’re interested in identifying the most expensive items in your online catalog, you’ll need to sort the data so those items appear at the top of the table.

These kinds of manipulations can be done with a programming language (e.g., R or Python) or within a spreadsheet application (e.g., Excel).

### Pivoting

*Pivoting* the data means turning or rotating it. In the most common use of pivoting, you restructure the data by transposing rows and columns: data in rows are moved into columns, and data in columns are moved into rows. You can do this to an entire table or just a portion of it (Figure 1).



*Figure 1: Data pivoting*

### Shaping

*Shaping* the data can mean selecting only one column or perhaps adding a new column derived from the values in one or more other columns. Suppose you have a column called *new patients* and a second column called *existing patients*. Now you want a new column called *total patients*. You would need to reshape your data to include this new column.

### Grouping

Data can be aggregated in many different ways (see Tables 1 and 2). For instance, you might apply mathematical operators to numerical fields or you might combine several descriptive fields together. Or you can do both. For example, suppose you have sales data for several stores and you want to group the sales figures by store zip codes, the total number of employees, and the store’s age (i.e., “old” or “new”). This involves manipulating the data.

| **Country (All)** | |
| --- | --- |
| **Row Labels** | **Sum of Order Amount** |
| Amy Dotson | 75048.04 |
| 7/15/2006 | 2490.5 |
| 7/31/2006 | 1873.8 |
| 10/10/2006 | 5275.71 |
| 10/21/2006 | 88.5 |
| 12/25/2006 | 166 |

*Table 1: List of dates before grouping*

| **Country (All)** | |
| --- | --- |
| **Row Labels** | **Sum of Order Amount** |
| Amy Dotson | 75048.04 |
| Q1 | |
| Jan | 6660.62 |
| Feb | 20418.34 |
| March | 5401.05 |
| Q2 | |
| April | 10881.61 |
| May | 555.6 |
| June | 3482.5 |

*Table 2: List of dates grouped by quarters and months*

# Data Visualization

Here are a set of simple steps for selecting appropriate data visualization:

**Step 1:** Identify your audience.

**Step 2:** Identify questions to be answered by your visualization.

**Step 3**: Think of possible constraints or challenges (e.g., budget limitations, the speed at which the study can be completed, or the dependability of your data) and ask the following questions:

* Is my data reliable?
* Is the value of the analysis greater than the cost?
* How will I store my data?
* Is the data owned or controlled by someone else?

**Step 4:** Select the appropriate visualization.

## Advantages and Limitations

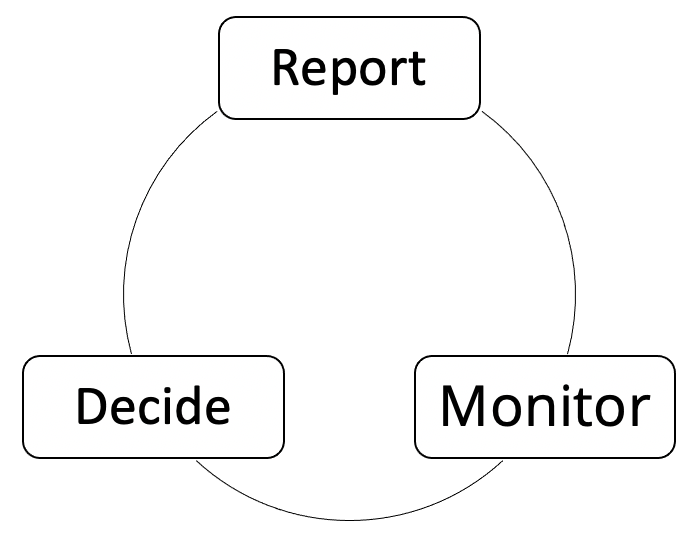
“A picture is worth a thousand words,” goes the old adage. The purpose of visualization is to communicate information quickly and efficiently. A chart plotting the distribution of some measured quantity—say the height of everyone in your hometown—certainly communicates the characteristics of that distribution more expeditiously than trying to interpret a set of means, standard deviations, skewness, kurtosis, etc., or examining a table listing each measurement separately.

However, visualization can be a double-edged sword. There’s a danger of “over-summarizing” data in a graphic simply by trying to display too much information. For example, suppose you want to show that your company’s products are selling better in one region of the country than in others. You can show revenue figures superimposed on a geographical map, but if you want to show additional information such as the names of the salespeople responsible for the most sales and the names of the top-selling products, the map risks becoming cluttered and difficult to read.

## Reporting, Monitoring, and Deciding

*Reporting* involves the presentation of tables, data, or information. *Monitoring* refers to looking at these reports consistently with an eye to making decisions or taking actions as needed to make business course corrections. This is effectively a feedback loop in which reports containing analytical results (the “what/why/who” of the organization) are monitored and then some actionable decisions are made (including a decision to take no action, if appropriate).

For example, if a report generated at a factory indicates that some of the machines are down, decision-makers monitoring the reports could respond by sending for a repair crew.



*Figure 2: The cycle of reporting, monitoring, and deciding*

## Tools and Techniques

Choosing the best graphic, such as a line chart, histogram, or map, depends on the data with which you’re working, your audience, and your tools. The simplest kind of visualization is often a chart.

### Charts

A chart is a graphical representation of data in the form of a bar chart, line chart, pie chart, and so on. For a measured quantity, you can use a chart to show magnitude (its impact), statistics (means, standard deviations, etc.), distribution, change over time, and in some situations flow. Charts are arguably one of the most popular graphic representations because they can be created with almost any tool ranging from programming languages (e.g., Python or R) to business intelligence tools (e.g., PowerBI or Tableau).

### Histograms

A histogram is a special type of chart that merits some additional explanation. To construct a histogram, one takes the range of possible values that your measured quantity can take and divides it into a fixed number of “bins.” Each bin represents a subset of the total range that the measured quantity can take. Finally, you count the number of measurements that fall in the range of each bin and plot the results. Returning to the example of people’s heights, suppose everyone’s height ranges from 48 inches to 84 inches. You would then divide this range into, say, six bins with each bin being six inches wide. The first bin ranges from 48 inches to 54 inches; the second bin ranges from 54 inches to 60 inches, and so forth all the way up to 84 inches. Next, you (or your data analysis software) would count the number of height measurements that fall into each bin and finally plot the results.

### Maps

Maps are very relevant in a world where events have geographic underpinnings. Data associated with geographic and geophysical phenomena (e.g., the extent of wildfires, locations of earthquakes, and weather patterns), business activity (e.g., sales activity by region), and politics (e.g., election results by state and county) are all amenable to being displayed on maps in revealing and compelling ways.

# Data Storytelling and Communicating Insights

After you have accumulated data, compiled it, organized it, sorted it, identified its business value, and developed a message that you want to communicate, it’s time to make a presentation to project stakeholders. This is where storytelling enters the spotlight. A compelling story provides a context that sets the stage for your observations and recommendations and allows stakeholders to make informed decisions about the project’s direction based on your findings.

Here are several pointers for effective storytelling:

1. Know your audience, know your audience, know your audience. Establish your goal or desired outcome for the presentation. Do you want to inform or make a push for action? What is the problem you’re trying to solve?
2. Prepare, process, and thoroughly understand your data. Produce your analytics or models, perform exploratory data analysis (EDA), and create some visualizations.
3. Identify the specific statements, comments, or messages that you wish to communicate.
4. Repeat Steps 2–3. Make sure you can find the holes in what you’re communicating and then refine your message. Focus on clarity.
5. Make the presentation to your audience.

Several simple rules of thumb:

* Know your context.
* Practice ethics (i.e., don’t manipulate or obfuscate data in a way that leads to misleading or false conclusions)
* Simple messages overshadow complex ones.
* Make sure all graphs, charts, or graphics have clear labels. Remember, simplicity is a virtue.

## Persuasion and Perspectives

As you work with data and extract business value from it, it’s critical to keep your audience in mind. For example, in a business setting, you must understand your audience’s perspectives as business owners, employers, and employees. Business owners may seek increased profits, employers may strive for increased employee productivity, and employees may look for a better work-life balance. Thus, the perspective you take in analyzing your data and making your presentation must match that of your audience. Effective persuasion depends on first finding relevant metrics that your audience understands and then focusing your attention on just those metrics.

As another example, suppose you’re working on a research project trying to answer the question, “Why are people getting sick when they shop in a particular store?” The information you need includes the type of illness, the dates and times when people who got sick were shopping in the store, the merchandise they purchased, and related relevant information. Daily sales figures for the store are *not* relevant as they won’t help you identify the source of the illness. Knowing the question that your analysis seeks to answer helps you identify the questions you need to ask when collecting data so you can focus on what’s relevant. Otherwise, you may find yourself presenting information and conclusions that are not applicable.

Finally, it’s important to understand your limitations. You might have a small team, a small budget, and a tight timeline, and you may not be able to produce the expected results. Understanding and communicating your limitations from the start can help you avoid this pitfall.

## Individual Visualizations and Dashboards

Sometimes, a single visualization is not enough to convey all the underlying information. Several related and often integrated graphs/charts can be used within a single screen to provide a more complete and cohesive visualization. In data analytics and business intelligence, these composite visualizations are usually called *dashboards*, so-named because they look like an automobile or airplane dashboard on which critical information is displayed using a variety of visualizations, dials, and gauges. These digital dashboards convey summaries and alerts in the form of graphs, charts, dials, and gauges that are tied to key performance indicators for the business. They are often an indispensable tool in business analytics.

## Ethical Data Analytics

Unfortunately, in business it is sometimes tempting to seek profit at the expense of ethics. But they don’t need to be mutually exclusive. When there’s a conflict between the two, it’s best to choose ethics over profit.

For instance, it’s possible to prey upon people’s health fears by selling unneeded medicines, supplements that don’t work, or equipment that provides no benefits. Though it may be profitable, it is unethical. Accordingly, it would be unethical to manipulate data in a way that shows those medicines, supplements, and equipment as indeed providing benefits. Manipulating data includes omitting “inconvenient” data points, adjusting visuals to mask or obscure features in the data, and of course outright falsification of data elements.