Analysis and Data Exploration

Introduction to Analyzing Data for Business Goals

# Organizing and Analyzing Data

By choosing the right data (and of course the right sources), you can develop business-relevant analytics and models that allow you to predict and optimize business outcomes. You can:

* Use data to improve your business
* Make data-driven, evidence-based business decisions
* Understand what your customers think and how they feel
* Identify the latest trends applicable to your business
* Improve internal operations and processes
* Stay compliant with data protection regulations

# Considerations when Analyzing Data

Your first challenge is identifying or defining the specific data that can help answer your questions, draw accurate conclusions, and enhance your decision-making process. Here are five steps in the data analysis process to consider:

1. Define your questions.
2. Set clear measurement priorities.
3. Decide what to measure.
4. Collect the data.
5. Analyze the data.

## Interpretation and Subjectivity

After thoroughly analyzing your data, it’s time to interpret your results. As you begin crafting your conclusions, keep in mind that it’s not possible to *prove* a hypothesis true, Instead, one must *fail to reject* that hypothesis.

As you interpret the results of your data, ask yourself these key questions:

* Does the data answer your original question? How?
* Does the data help you defend against any objections? How?
* Are there any limitations on your conclusions or any angles you haven’t considered?
* Are the results more subjective than quantitative? This would mean that the information you’re dealing with is intangible and inexact, and thus difficult to collect and measure.

# Statistical Analysis vs. Data Analysis

What is the difference between *statistical analysis* and *data analysis*?

Remember, statistics and analytics are two branches of data science that are frequently used together. In short, analytics helps you develop hypotheses while statistics help you test those hypotheses.

*Statistical analysis* allows you to gain an understanding of a population from an analysis of a small but representative subset (a sample) of that population. In turn, *data analysis* is a procedure for investigating, cleaning, transforming, and training the data with the aim of finding useful information, drawing conclusions, and helping with decision-making (Figure 1). Data analysis tools include Open Refine, Tableau public, KNIME, Google Fusion Tables, Node XL, and many more.

To summarize, data analytics involves the use of data, machine learning, statistical analyses, and computer-based models to derive better insights and make better decisions based on the data.

| **Statistical Analysis** | **Data Analysis** |
| --- | --- |
|  |  |

*Figure 1: A comparison of statistical analysis and data analysis*

# Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) employs a variety of techniques (mostly graphical) to:

* Maximize insight into a data set
* Uncover underlying structure
* Extract important variables
* Detect outliers and anomalies
* Test underlying assumptions
* Develop models
* Determine optimal factor values

The main purpose of EDA is to characterize the data and produce a variety of descriptive statistics such as averages and medians. As you might expect, more sophisticated questions about the data require significantly more work in preparing the data. Typical objectives for EDA include identifying missing values and problematic units and discovering data that requires some sort of transformation before it can be analyzed.

## EDA Techniques

EDA encompasses both non-graphical and graphical techniques. Most techniques in use are graphical, but there are also several purely quantitative techniques used, as well. Here are a few examples:

* **Plotting the raw data**: Data traces, histograms, probability plots, lag plots, and block plots
* **Plotting simple statistics:** Mean plots, standard deviation plots, box plots, and main effects plots
* **Positioning plots:** Maximize natural patterns by using multiple plots per page

Several particularly popular plots include:

* Stem-and-leaf display
* Resistant time series smoothing
* Scatterplot smoothing
* Median polish
* Bubble chart
* Resistant curve fitting
* Multi-vari chart
* Violin plot, wind rose
* Diamond plot
* Heat map
* Population pyramid
* Sunflower plot

Figure 2 shows a sample of templates from the application Tableau including several popular EDA plots:



*Figure 2: Templates from Tableau*

Essentially any analytical method that looks at a dataset falls under the category of EDA. In addition to simply understanding a dataset, EDA helps you check assumptions, look for data errors, generally characterize the data, examine relationships that may exist among the variables, and consider how the data might be used in models.

EDA begins with a combination of input data and assumptions collected from a particular audience or set of individuals who have specific questions or needs associated with the data. The data then goes through a process of cleansing, preparation, and transformation. Needless to say, some datasets are in better condition than others, and a particular dataset may need to be put through this process several times, improving in quality each time.

## Characterizing the Data

Data characterization involves summarizing the distinctive features of a dataset such as the meaning of the fields, the number of columns and rows, and whether it is structured or unstructured. The dataset is effectively “revealed” when characterized visually and statistically.

### Outliers, Nulls, and N/As

One important objective of data characterization is identifying data elements that are nulls or voids—which are effectively the absence of any element. These are not zeroes or empty strings—they are truly an absence of data. Sometimes we code nulls or voids with numerical data values such as “99999” or “missing.”

## Understanding the Data

*Understanding* data comes after or sometimes at the same time as characterizing it. “Understanding” means knowing and being able to explain what the data means. For example, suppose you’re working with financial data or medical data. What does a column labeled “INS” mean? Is it “insurance” or do you need a data dictionary to define it? Perhaps you can look at the values in the column and figure out what the column means. If you know that there are multiple service providers, for instance, and there are additional columns such as “COST” in the dataset, you might conclude that “INS” and “COST” are related, such as individual providers having certain associated costs.

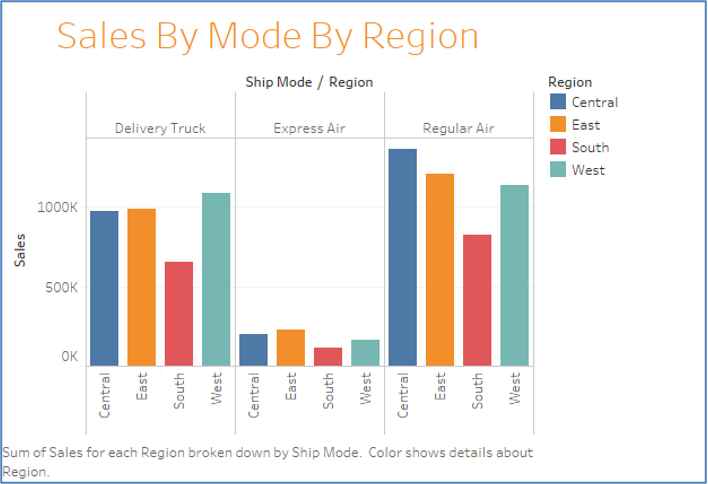
### Graphs and Descriptive Tables

As alluded to earlier, data often comes in a *structured* format, in columns and rows (recall the columns labeled “INS” and “COST”). This data can then be presented and stored electronically as a table with a defined number of rows and columns. This structure is also referred to as a table’s *schema*. For example, a table might contain patient health records or information about countries participating in humanitarian efforts. In contrast, *unstructured* data consists of the contents of documents (text), images, or videos—anything that is normally not amenable to being organized as a table.

Understanding data implies that you’ve cleansed it, prepared it, and transformed it, and likely vetted it with stakeholders. At this point, you are intimately close to it and are able to present it to a given audience. Perhaps the best way to both understand and present data is visually through graphics. Below are a few examples of how data can be presented both in a table and via graphics (Figures 3 and 4).



*Figure 3: An example of how data can be presented in a table*



*Figure 4: An example of how data can be presented in a graph*

## Variables

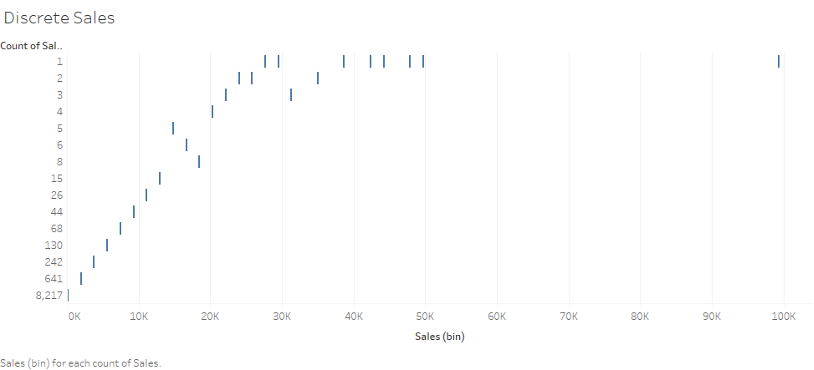
### Quantitative and Qualitative (Categorical)

* **Variable:** A measurable or observable quantity that varies from one person or thing to another
* **Qualitative variable:** A non-numerical variable
* **Quantitative variable:** A numerical variable

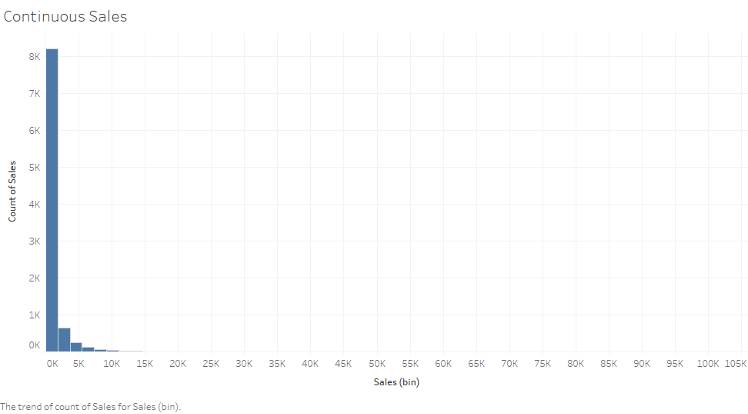
### Discrete and Continuous

* **Discrete variable:** A quantitative variable whose possible values can be listed individually—in particular, a quantitative variable with only a finite number of possible values is a discrete variable (Figure 5)
* **Continuous variable:** A quantitative variable that can take any numerical value over a given interval (Figure 6)

Figure 7 shows how (1) a variable can be divided into qualitative and quantitative components, and (2) quantitative variables can be further divided into discrete and continuous values. Table 1 shows several characteristics for continuous and discrete variables.



*Figure 5: Graphic showing a discrete variable*



*Figure 6: Graphic showing a continuous variable*



*Figure 7: Schematic relationship among discrete, continuous, quantitative, and qualitative variables*

| **Discrete Variables** | **Continuous Variables** |
| --- | --- |
| * Have only a finite or countably infinite set of values * Include zip codes, counts, or the number of words in a collection of documents * Are often represented as integers | * Have real number values * Include such quantities as temperature, height, or weight * Can only be measured and represented using a finite number of digits * Are typically represented as floating-point variables |

*Table 1: Several characteristics of discrete and continuous variables*

## Data Types and Plotting

In computer science, a data element is categorized in terms of its *type.* Just as an apple is a *type* of fruit and a cat is a *type* of animal, data also falls into broad categories. The most common data types include:

* *Floating point* (e.g., a number)
* *Boolean* (e.g., having one of two possible values such as true/false or yes/no)
* *Character* (e.g., non-numeric value, such as a letter of the alphabet)
* *String* (e.g., a series of characters)
* *Integer* (e.g., non-fractional numbers)

Different data types are treated differently when processed by a computer program. For example, if you’re dealing with money in a bank account, the values for credits, debits, and balances would be *floating point* numbers. You can add, subtract, multiply, and divide these numbers. If you’re dealing with the names and addresses of account holders, you would use strings consisting of a series of characters. You can print these strings on address labels. If you’re storing the account holders’ ages, you would use *integer* values (whole numbers). Finally, if you’re keeping track of whether they prefer to receive their monthly statements by mail or electronically, you would use a *Boolean* value since there are just two possibilities. Needless to say, it doesn’t make sense to add a string to a floating point value or divide a character by an integer.

Paying attention to data types helps:

1. Avoid programming mistakes such as trying to add an integer to a character.
2. Reduce the amount of electronic storage needed to save your data.

This second point involves the cost of data storage, which can be quite high when you’re dealing with many data fields for potentially millions of customers. Floating point values use the largest amount of space while Booleans require the least (because they have only two possible values). For instance, in the interest of saving space (and money), it’s not a good idea to use a floating point value to store a person’s age (integer).

For readers interested in additional classifications of these data types, refer to Figure 8, below.



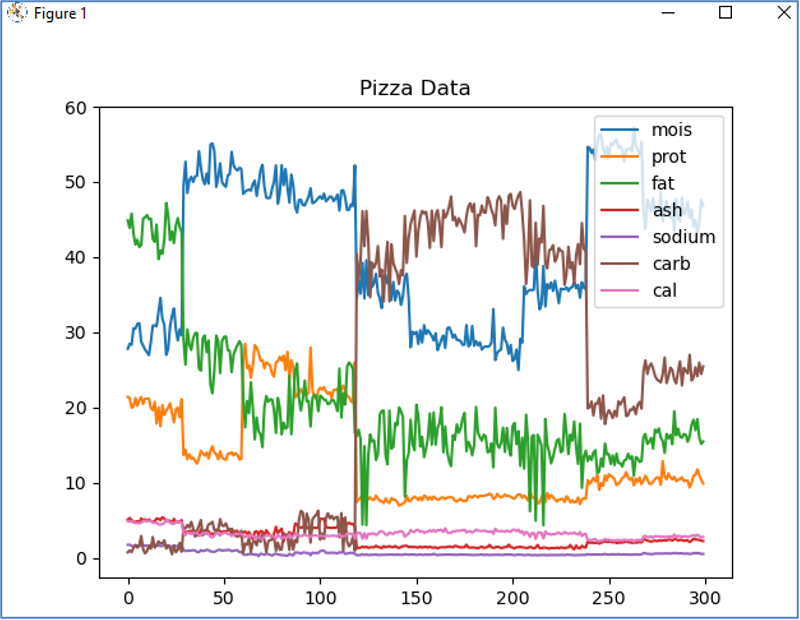
*Figure 8: Data type classifications*

Keep in mind that different programming languages may treat the various data types in slightly different ways. For more information, you can consult books and videos on the specific programming language of interest.

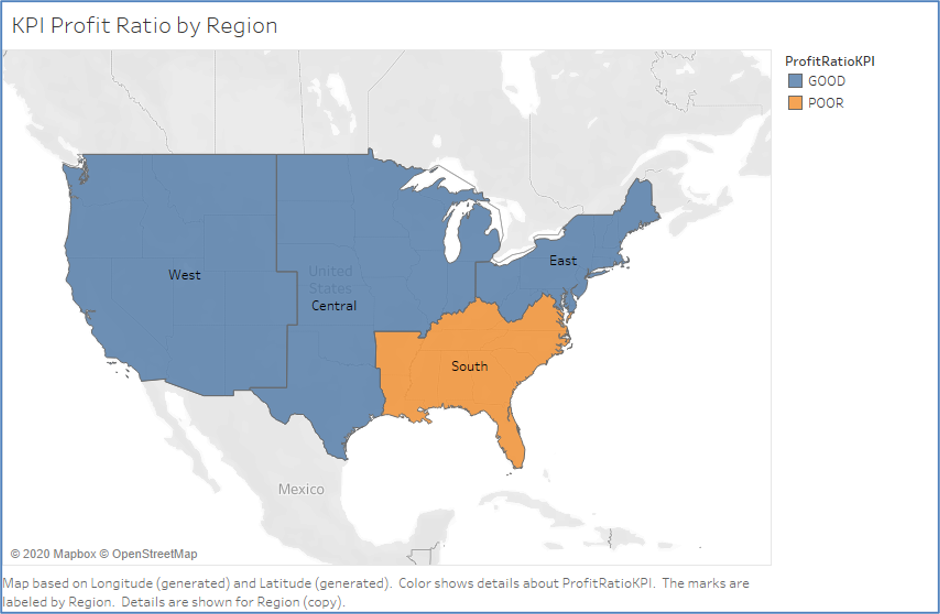
### Data Plotting

Should you plot your data? It depends! The purpose of plotting data is to quickly gain a sense of variations and relationships that may be present. But not all datasets require a plot. If there are only a few points, you can examine the numbers directly since a plot won’t necessarily reveal any additional information. Similarly, if there is minimal variation in the data, it is easy enough to see or state the fact without using a graph of any sort.

We won’t go into detail here on the many different kinds of plots that analysts commonly use, but just to give you an idea, take a look at the two plots that appear below (Figures 9 and 10). The first one shows seven laboratory measurements where the y-axis values vary continuously as a function of the x-axis values. The second plot shows whether a “profit ratio KPI” is “good” or “poor” for different regions of the country. The analyst could have used a simple table showing “good” or “poor” for each region, but using an actual map makes the information more attractive and compelling. “Good” and “poor” comprise an integer or Boolean data type, though if there were more categories such as “superior” or “unacceptable,” you would need to use an integer data type to represent all of them.



*Figure 9: Data presented as a set of line graphs*



*Figure 10: Data presented as a map*

## Data Cleaning and Maintenance

### The Benefits of Clean Data

When using data in a predictive model, it’s important to ensure the data is “clean.” Ensuring that it’s accurate and free from errors—at least as best as possible—gives you confidence that your modeling results will be meaningful and will allow you to draw valuable conclusions. Think of a predictive model as a simple system in which the data comprise the input values and the model’s results comprise the output values. “Clean” data will give you equally “clean” results.

### Data Maintenance

Many organizations use datasets that change frequently as new information is added and outdated information is culled. This requires a continual process for reviewing the data, verifying its accuracy, and correcting it so the organization can have confidence in any analytical results produced using the data. Data checks should be performed at regular intervals—daily or weekly, for instance.

### Common Tools and Best Practices

A number of commercially available tools for cleaning and maintaining data exist. They range from the simple, such as Microsoft Excel or similar spreadsheet applications, to the complex, such as an enterprise-scale system. Several of the more commonly used data analytics tools that have data cleansing capabilities include Python and its data analysis library, pandas, Google Refine, DataWrangler, Datamartist, MoData, AnalyticsCanvass, TIBCO Clarity, SAS Data Management, Tableau Prep, Data Ladder, and IBM InfoSphere.

As a best practice, it’s good to have a data cleaning plan as follows:

* Develop a data quality plan with set expectations for your data
* Validate the accuracy of your data (in real-time, if possible)
* Identify duplicates
* Append data as needed
* Perform regular maintenance
* Conduct regular backups

# References

* Glen, S. (2020). [Difference between Data Analysis and Statistical Analysis.](https://www.datasciencecentral.com/profiles/blogs/difference-between-data-analysis-and-statistical-analysis)
* State Graphics 19. (n.d.). [Exploratory Data Analysis.](https://www.statgraphics.com/exploratory-data-analysis)