Introduction to Descriptive Analytics

Introduction to Analyzing Data for Business Goals

# What is Descriptive Analytics?

In a business setting, descriptive analytics is the process of analyzing historical data with the goal of examining changes in the business’ operations and outcomes. For example, a company may track its performance over time by recording monthly sales figures and comparing them to previous years. Based on the trends it observes, the company might then make changes to its operational strategy.

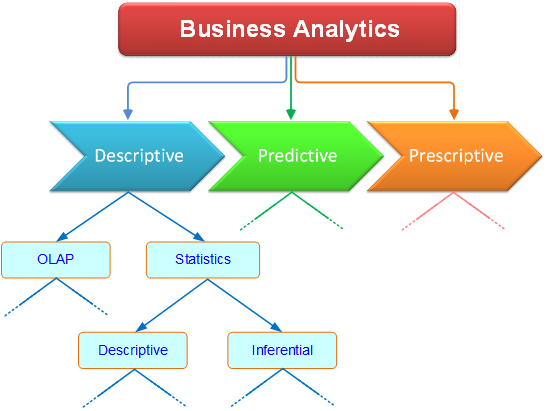
Descriptive analytics comprises the earliest part of the analytics chain, preceding diagnostic, predictive, and prescriptive analytics. It sets the stage for being able to make diagnoses, predictions, and recommendations for specific actions in support of business goals. It uses a variety of tools—including descriptive *statistics*—to condense and summarize data to foster an understanding of the changes the business may be undergoing.

Let’s look at social media as an example. Twitter stores an immense amount of data in the form of tweets. Let’s say we want to find the total number of tweets related to the National Association for Stock Car Auto Racing. First, we need to consider any synonyms or abbreviations for this organization (i.e., car racing, fast cars, or NASCAR). Next, we need to determine how the data is stored on the Twitter platform and how it can be accessed, and then identify the various steps needed to extract, store, and analyze it.

As discussed in previous modules, the process involves several steps including data aggregation, mining, and preparation. However, for the purposes of this example, let’s assume that we've already completed all these steps and have a dataset that is ready for analysis. For instance, we can apply simple filters and find states that have residents interested in NASCAR (say, North Carolina, California, or Kentucky). If we wanted to understand the NASCAR demographic, we could group the people posting tweets about NASCAR by age or gender (Twitter might charge us for this data if it is even available).

## Techniques of Descriptive Analytics

Generally speaking, descriptive analytics has two main branches: statistics and online analytical processing (OLAP). OLAP is the term used for analyzing, characterizing, and summarizing structured data stored in organizational databases (often stored in a data warehouse or in a data mart) using cubes (i.e., multidimensional data structures that are created to extract a subset of values to answer a specific business question). The OLAP branch of descriptive analytics has also been called *business intelligence*. Many popular data visualization tools such as Tableau, PowerBI, Spotfire, etc. use an automated multi-dimensional data cube generation process (called VizQL) to create informative graphical illustrations.



*Figure 1: Branches of descriptive analytics*

Statistics, on the other hand, helps characterize data in terms of individual variables or multiple variables by using either descriptive or inferential methods.

### Descriptive Statistics

Descriptive statistics numerically describe the essential characteristics of a set of data, giving you a sense of the data with which you’re working. Because descriptive statistics are calculated using existing data, the implication is that they describe events or measurements that have taken place in the past. In contrast, *inferential statistics* are used to draw conclusions that are broader in nature, such as using survey results to infer people’s attitudes toward a particular issue.

The descriptive statistics used to summarize individual data series are usually grouped into two categories: centrality (mean, median, and mode) and dispersion (standard deviation, variance, and range). The most common descriptive statistics are:

**Mean:** Also known as the *average*, the mean is calculated by summing the values in a dataset and then dividing by the number of values. The mean provides a notion of centrality in the data.

**Median:** This is the value in the middle of the given numerical data series. It is determined by sorting the data and identifying the specific value of the data point in the middle. If the data series has an even number of instances, then the two data points in the middle are averaged to calculate the median.

**Mode:** The mode is the value appearing most frequently in a dataset.

**Standard deviation:** The standard deviation measures the amount of variability among the values in a dataset.

**Variance:** The variance is the square of the standard deviation.

**Maximum and Minimum:** These are the largest and smallest values, respectively, in a given numerical data series.

**Range:** The range is the numerical distance between the maximum and minimum values.

**Correlation:** The correlation is a measure of the linear relationship between two variables.

Let's say we want to analyze patient data from several hospitals. Descriptive statistics would give us the answers to the following questions:

* How many patients are in each hospital?
* What is the average spending of patients in each hospital?
* What is the median spending of patients in each hospital?
* Which type of patient most frequently returns to the hospitals?
* What is the standard deviation of patient spending in each hospital?

### Summary of Data

Let’s go through an example of how descriptive statistics are used to summarize a dataset. Suppose you work for a company that manufactures light bulbs and you want to analyze how long the bulbs last before they burn out. You test 20 bulbs and find the results shown in Table 1, below.

|  |  |
| --- | --- |
| **Bulb No.** | **Time (Hours)** |
| 1 | 669 |
| 2 | 630 |
| 3 | 745 |
| 4 | 598 |
| 5 | 615 |
| 6 | 610 |
| 7 | 653 |
| 8 | 733 |
| 9 | 801 |
| 10 | 742 |
| 11 | 789 |
| 12 | 607 |
| 13 | 632 |
| 14 | 692 |
| 15 | 607 |
| 16 | 806 |
| 17 | 753 |
| 18 | 695 |
| 19 | 779 |
| 20 | 717 |

*Table 1: Light bulb test results*

These numbers give you a general sense of the expected longevity of the bulbs. However, if you want a concise summary of the results, you need to calculate some descriptive statistics:

Count: 20 (the number of light bulbs tested)

Mean: 693.65 hours (the average time the bulbs lasted)

Standard Deviation: 72.03 hours (the variability of the measured times)

Low Value: 598 hours (the bulb that burned out the soonest)

High Value: 806 hours (the bulb that lasted the longest)

Mode: 607 hours (the number of hours that *most* of the bulbs lasted)

There are a variety of software applications such as Microsoft Excel that can easily calculate these values. The results give you a summary of your entire dataset in only five numbers. Suppose you had tested 1,000 light bulbs. You would still be able to summarize those results with just five numbers.

These statistics become more useful if you use them to compare different samples. Suppose you test a sample of 1,000 light bulbs made with a standard filament and another sample of 1,000 made with a new-and-improved filament. Calculating descriptive statistics allows you to quickly compare the longevity of the two samples.

# Considerations for Creating Descriptive Analytics

## Asking the Right Questions

One of the first things you must do when starting a research project or tackling a business problem is to identify your audience’s objectives and develop a corresponding problem statement. This shapes the information you need to collect in order to conduct your analyses. Your audience will often be able to provide you with this information, but sometimes you need to “tease” it out by asking the “right” questions.

Suppose you’re hired to conduct some descriptive analytics for a customer in the airline industry. They’re interested in examining how the most recent global pandemic has affected their airline’s business. The descriptive statistics you wish to calculate thus form the questions you need to ask:

* How many passengers were flying before and after the pandemic? [count]
* What was the average number of passengers flying before and after the pandemic? [mean]
* Who are the most frequent flyers? [mode]
* What was the variability of passengers flying before and after the pandemic? [standard deviation]

There are several additional questions worth asking but more difficult to answer:

* What are the expectations for a vaccine?
* What is the optimal seating capacity that maximizes both safety (in terms of physical distance between passengers) and profitability?
* Can we pre-purchase fuel before the price of oil (and jet fuel) increases as expected?

## Selecting Valuable Data Sources

Once you understand your audience’s needs, the problem statement, and the questions to be answered, you can begin searching for and choosing data sources. Useful data comes in a wide variety of formats. For instance, it can be structured (as in a formal relational database) or unstructured (as in a set of files on a drive).

Several considerations in the search for data include:

* **Cost:** How much will this data cost me or my organization? Does the value it provides justify the cost?
* **Risk:** What are the potential errors in this data and what transformations are needed? What are the data’s limitations? Does it go back in time far enough? Does it contain all the information needed? Does my organization’s competition rely upon the same data (this might erode the data’s competitive value)?
* **Return:** Will the data ultimately lead to increased product sales, better customer retention, or reduced staff turnover? Is the dataset unique (meaning that it potentially can provide my organization with a competitive advantage)? Can I process the data in a unique way that gives my organization an “edge”? Can I customize the data to “fit” my organization’s specific needs? Can my organization store the data and use it to its advantage in the future?

While there are no “right” or “wrong” answers to these questions, the ability to find answers to most of them suggests that the data source you’re considering probably has value.

# Business Challenges

## Missing Time Series Data

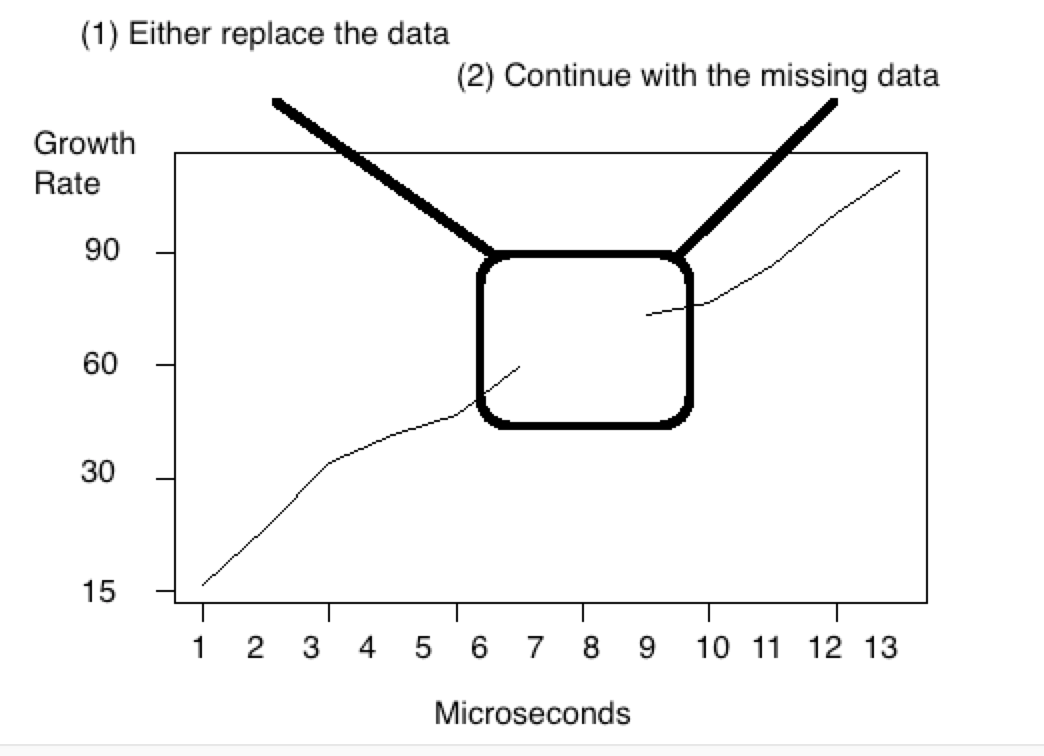
A *time series* consists of a set of data points—measurements or other data—that occur at regularly spaced time intervals. For example, you can measure the air temperature once per hour and plot it on a chart. These measurements constitute a time series.

Very often, you have to deal with time series that have *missing data*, which might be the result of someone forgetting to check the thermometer at one of the appointed times. Also, in some situations, you know in advance that a value should be missing. For instance, suppose you’re tracking daily sales figures for a store that is closed on Sundays. When you do an analysis of those figures, you have to account for the fact that Sunday data will *always* be missing! In general, if you neglect missing data or if you don’t understand why a data point is missing, you might draw conclusions that don’t reflect reality.

There are two solutions to missing time series data:

* Use mathematical interpolation to estimate the missing value based on adjacent values in the series
* Remove the data point completely

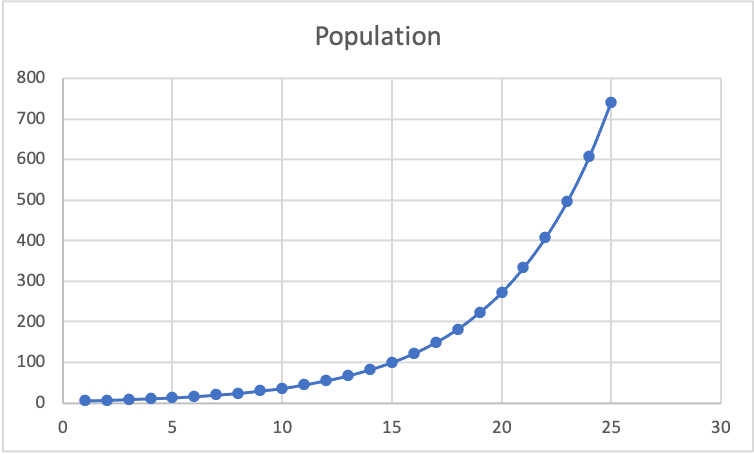
For example, if a store is closed on a holiday, you might estimate the sales figure for that day by averaging the sales figures for the day before and the day after. This makes your time series continuous and is a common practice in data analysis for ensuring an unbroken time series. Alternatively, you can omit the sales figure for the holiday altogether but ensure that any calculations you perform are not biased because of that missing data point.



*Figure 2: Time series with missing data*

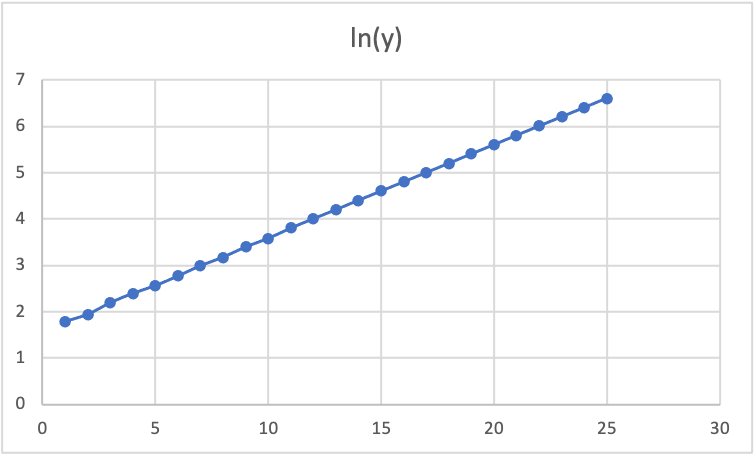
## Transforming Variables

Suppose you’re doing biological research and are tracking the replication of a virus within a host cell. You plot the number of individual viruses as a function of time (in minutes), which results in the graph shown in Figure 2.



*Figure 2: Virus population growth as a function of time*

This looks like exponential growth, which is what you would expect from a virus. However, if you apply the logarithm function to each value in the chart, you’ll get a linear plot as shown below.



*Figure 3: Virus population growth as a function of time plotted on a semi-log graph*

Now you see a straight line, which confirms the exponential nature of the virus’ growth. Any deviation from what you would expect shows up more easily when looking at what is supposed to be a straight line than when looking at a curve. Taking the logarithm of each value is called a *variable* *transformation* (in particular, this is a *logarithmic transformation*). There are many different ways you can transform your data—your variables—into values that are easier to read or understand.

Here are three examples of common transformations:

* **Rescaling a variable**: Suppose you keep track of local temperatures in degrees Fahrenheit but your friend keeps track of temperatures in *her* hometown in degrees Celsius. Before comparing temperatures, you would need to convert her measurements into degrees Fahrenheit (or yours into Celsius). This is one type of rescaling.
* **Normalizing a variable:** Normalizing a variable means adjusting its values so they all lie between zero and one but still maintain their relative relationships. This shouldn’t be confused with *standardization*, in which a normally distributed variable has its values adjusted so the mean is zero and the standard deviation is one. (Standardization is an advanced topic from statistics.)
* **Label encoding:** Suppose you’re analyzing the results of a multiple-choice test in which the answers are categorically labeled a, b, c, d, and e. You can transform this categorical variable into a numerical variable by assigning a=1, b=2, c=3, d=4, and e=5. Such a transformation may be necessary if you’re applying a computational algorithm to the data.

## Interpretation of Meaning

Even if you have the most sophisticated analytical software at your disposal, the quality and meaningfulness of your results depend on how you interpret those results. Let’s look at a very simple example.

As an official in the county’s transportation department, you want to know how commute times have changed as a result of the COVID-19 pandemic. You hire a researcher to help you find out.

Before the pandemic, most people were going “out into the world” to their workplaces, but after the pandemic started, most people stayed home. Your researcher decided to survey a large number of commuters randomly chosen from the local population and asked them to report their typical weekly commuting times for a five-year period before the pandemic and their typical times after the pandemic struck. Here are the average driving times per person per week:

Before the pandemic: 40 hours

After the pandemic: 5 hours

Average (before and after) = (40+5) / 2 = 22.5 hours spent commuting

Thus, your researcher happily informs you that the average commuting time is 22.5 hours per person per week.

But does this make any sense? Remember, the 40-hour figure represents a much longer timeframe (five years) than the 5-hour figure and probably should be weighted accordingly. Looking at the previous six years (five years pre-pandemic and one-year post-pandemic), it makes sense for the average time to be closer to 40 hours. However, with the pandemic in full swing, the average time is indeed just five hours. So, averaging the two averages doesn’t produce a meaningful result. The more useful piece of information is the difference between pre-pandemic and post-pandemic commuting times, which dropped from 40 to five hours. This is the information that will help you make policy decisions.

**Group Discussion:**

Reading the marketing science journal and replying to the following questions”: The Value of Descriptive Analytics: Evidence from Online Retailers by [Ron Berman](https://www.researchgate.net/profile/Ron-Berman) and [Ayelet Israeli](https://www.researchgate.net/profile/Ayelet-Israeli).

1. What are the differences between the descriptive and inferential analysis in the paper? Can you find any examples?
2. What the mechanism or metrics this paper used to measure and guide the e-commercial retail marketing? Find those methods and compare to the any other mechanism you have used before (e.g., Google analytics)
3. (Optional) How the analytics results are used in the marketing action? Do you agree or disagree with the conclusion and why?