Data Essentials

Categorizing Data

Varieties of Data

Data can come in many forms, not just numbers, which requires us to have a flexible view of what constitutes data. One approach is to think of data as a recording of the world around us. With this definition, text descriptions, sound recordings, video, and pictures all qualify as data. Considering these items to be data can get complicated when we begin to analyze those datapoints. The more traditional analysis techniques were created for numeric values and apply statistical methodologies but that won't work with text, audio, or images. How would you find the square root of a word, or find the average of a picture? These non-numeric values need a different kind of analysis.

For this reason, data is often classified as quantitative (numeric data) or qualitative (text data, typically). Quantitative data is what most people think of when they think of data. These are the "measurables" of the world, such as weight, height, age, etc. These numeric values are easy to compare and fit nicely with traditional statistical analysis techniques. Numeric values lend themselves well to ranking in order, finding averages, and comparing across units.

Qualitative values use text to describe phenomena in the world, such as hair color, eye color, etc. Often these values are associated with categories (aka categorical data) which allow for grouping of information. Those groups can be ordered (ordinal) like below average, average, above average; or they can be independent like red, green, and blue. There are other times when qualitative data isn't meant to be categorical. Text documents contain information, and we need to extract that information from the text via a process known as text mining. This can involve looking for words that occur most often (topic modeling) or determining how the author felt by analyzing the language used (sentiment analysis). Working with qualitative data comes with unique challenges but it opens a lot of different possibilities for analysis.

At this point, you may remember that we mentioned videos, audio clips, and photographs as data, but we didn't mention them in qualitative vs quantitative discussion. Good catch! Describing data as qualitative or quantitative isn't the only way to categorize data. Another method is to describe the structure of the data, which refers to its format. This method categorizes data as either structured, semi-structured, or unstructured.

Structured data follows a regular format that is predictable, which makes it easy to know the types of values that are contained within. Structured data is commonly tabular data, which arranges values in tables. Columns contain values for a specific trait (like height or eye color). Each row of the table is associated with an observation. This means that all the values on a given row are associated with the same event (whether that be an individual or point in time). Structured data is the easiest to use for analysis because it has already been organized in a meaningful way that is consistent.

Data Essentials

Semi-structured data refers to data in formats that are regular but not necessarily standardized. Think about adding items to a cart in an online store. The items and quantities you select are data that can be stored in a tabular form but creating a table with every item offered by the store for each customer's cart doesn't make sense. Instead, they likely have some key information about you and the items you have selected. This data is convenient to store but requires some extra work when analyzing. (image of json with shopping items)

Unstructured data is everything else. This can be text documents, social media posts, audio clips, videos, and pictures. All of these items can be analyzed but they need to be organized before that can happen. Unstructured data accounts for most data collected (check your downloads folder).

File Types

File Types

As you continue working with data, you will need to collect it from many different sources, which means you will likely encounter many different file types. Some of these may be familiar to you already but others may be a little more abstract/specific. Let's start with the basics of why file types matter, how to determine the file type, and then review some commonly used file types.

File types are important because they determine how the data is stored and also help the computer to determine which programs should be used to open the files. You can find the file type of a file by looking at the extension behind the file's name (for example filename.file type). You can also view more about the file types by right clicking on a file and selecting "properties" in a windows environment or "get info" in macOS.

Let's take a look at some of the most commonly used file types and their extensions.

Spreadsheets:

Spreadsheets are commonly used to hold tabular data and perform basic calculations quickly. Each value is contained in a cell which can be identified by column and row. These files support formulas which allow you to perform calculations or manipulate your data. They are also called workbooks and each workbook can contain multiple sheets, which are navigated via the tabs in the interface. Common programs used for working in spreadsheets are Excel, Numbers, or Google Sheets.

- .CSV stands for comma separated values and is one of the most commonly used formats. This file type works with all of the spreadsheet programs listed above but only supports basic formatting of values.
- .XLS is a Microsoft Excel specific file type which supports formatting within that program. You may also see different version of this file type that end in x or m, indicating a particular format supporting specialized functionality.

Data Essentials

Documents:

Text documents are some of the most common file types, generally containing written text with support formatting options to stylize the text. Occasionally these documents will contain images or tables to supplement the text. These documents are generally created and edited in word processing programs such as Word and Pages, but they can also be created by simple text editors such as Notepad.

- .doc is the Microsoft Word file type and one of the most common document data types. Most editors can open .doc files but may require that you save in another file format.
- .txt is a plain text file which does not support formatted text or supplemental media.
- .pages is an Apple proprietary file type that has limited compatibility with other file editors. Exporting files to another format is necessary to make them accessible.
- .pdf stands for portable document format and is a popular option for saving finalized versions of documents that you intend to share. These files are difficult to edit without specialized software but are easy to open (and compatible with internet browsers) making them a good option for sharing.

Images:

Images are becoming more commonly used as data sources thanks to the emergence of cameras built into devices and surveillance equipment. Some of these images primarily contain text information, like scanned document or a picture of a restaurant menu. Other images don't contain text but may be analyzed and classified to determine changes or anomalies in physical characteristics. Image files can also be used to enhance reports or marketing materials.

- .jpg files are compressed image files which allow for simple tradeoff between quality & file size.
- .png stands for portable network graphic and is a popular file type for high quality images.
- .pdf can also be considered an image file type as it is often intended as a read-only format. Pdf files are commonly a source of text data in OCR (optical character recognition) workflows,

Audio:

Audio file types are more specialized within data analysis but they are becoming more common as they are often stored for historical record, and they are even being utilized as sources for machine learning transcription. The file type alone will not indicate the overall quality of an audio file because there are many factors that impact the quality. Still, knowing the most common file extensions can help you to easily identify audio data.

- .mp3 is one of the most recognizable audio file types thanks to mp3 players from the 2000s and is commonly used.

Data Essentials

- .wav is associated with high quality audio recordings because they don't compress the audio data as much as other file types. With improved quality comes increased file size and these files can be much larger than other options.

Video:

Video files are becoming more common with high quality video recording on phones, popular social platforms like YouTube and TikTok, and recordings of virtual meetings on Zoom. These files also tend to be the largest as they typically contain both audio and video data. Specialized software is required for editing & exporting video files.

- .mp4 is a standard file type that sometimes contains additional data like subtitles, in addition to the audio and video data.
- .avi is a file type specific to windows but can be played with most medial players.
- .mov is associated with the program Quicktime, which was developed by Apple.

Others:

The file types covered in this lesson are only a small subset of some common file types, but there are many others which you may encounter as you use specialized software in business environments. Software vendors will often have proprietary file types which include additional information specifically intended for their software. Outside of these exclusive file types, there are a few others that you'll likely run across which aren't program or file specific.

- .zip is a compressed file which is often used to send multiple files within a folder structure. The compression also removes redundancies between files by saving a single version of the items that are the same across multiple files. This fileype reduces the size of the files which is handy for sharing, but they do require the recipient to un-zip (or extract) the files before using them.
- .xml stands for "extensible markup language" and includes the ability to provide additional information along with data. Xml is an example of semi-structured data.
- .json stands for "javascript object notation" and is a format that uses "keys" paired with "values" to present data in hierarchical families. JSON is an example of semi structured data.

Data Types

Data Types

When trying to understand the data within any given file type, there are also data types to consider. Data types allow the computer to make assumptions about the data stored in the file. For instance, think of the numeric characters 9 0 2 1 0. A data type can tell a computer to treat these characters as the value 90,210, the date September 2, 2010, or to treat them as the zip code 90210 in California. It wouldn't make sense to try to add or multiply dates and zip codes but without data types, the computer can't make assumptions about the meaning of the value.

Data Essentials

They also help to define the expected format of a value, which can mean they restrict the length or type of characters that can be used. This is important for standardizing values and reducing the amount of space required to save information.

There are a number of data types commonly used but they are often heavily dependent on the file type/environment you are using. For instance, you can find many data types in Excel when formatting cells, many of which can also be found in Google Sheets. There are other data types used in coding languages like Python or Java. In general, you'll find that even if they aren't called the same thing, different programs or environments will use similar groupings of data types.

Numeric Data Types

As you might expect, numeric data types involve a bit of math. These data values contain primarily numeric characters and can be used to perform mathematical operations. Within numeric data types there are a few subcategories with different uses.

The integer family of data types is used for whole number values, like the quantity of an item or numbers that are rounded, such as age. There are different versions of integer data types with byte being the smallest of them. The rest of the data types in this family begin with INT followed by a numeric value. These data types differ in the range of values they can accommodate (which is based on the way computers store and process values as binary).

But what about calculations involving decimal values like monetary transactions or averages? For non-whole numbers, there are other data types that support storage and computation of decimal values. The names will vary depending on the environment you are using but some common names are float, double, and decimal. You may find that your environment offers multiple decimal compatible options to let you choose how specific the calculations and stored values should be. With these options, you can determine how ok you are with a little "rounding error" or if you need the most precise calculations with potentially reduced performance.

Keep in mind that just because you see a numeric character does not mean the data type is numeric. Things like credit card numbers and phone numbers are numeric characters that aren't suited to numeric data types. These values are unique identifiers and don't have value relative to one another. For instance, a phone number of 867-5309 isn't better or worse than 867-5308, they just represent different codes. There are also occasions where you will want to include non-numeric symbols when formatting columns, like currency or percentage, which may require converting values to non-numeric data types.

String Data Types

String data types are used to store text values, and while that may seem straightforward, there are a number of subtypes in the string category that specialize in accommodating a variety of needs. We will consider two common factors that come up when working with string data.

Data Essentials

One of the factors that is useful to consider when choosing a string data type is the length (or number of characters) that comprise the data values. While you may not know exactly how many characters there are in each value, in some cases it is common to set a limit on the number of characters that can make up any one value. This fixed-length option will reserve the same number of characters for every entry, making them predictable but not always efficient. Sometimes this strict limit can be an issue so you may opt for a data type that is flexible, commonly referred to as a variable-length.

Another factor to consider is the ability to accommodate "special" characters which are commonly required for languages other than English. To understand why, let's consider what happens when you enter characters into a computer. Regardless of how you enter the information, via a keyboard, touchscreen, or even speaking to a voice assistant, those characters end up as ones and zeros in the computer. This is known as encoding characters, and it requires a standard that defines how each character should be encoded. There have been different standards developed for different languages, for instance ASCII was commonly used for English speakers and the accompanying characters set. As the world has become more global, we often need to accommodate a larger set of characters for other languages which has led to the adoption of more recent standards like Unicode. Some data types only accept the older ASCII or Latin characters while others will support Unicode or Wide string values. If your data values do not display correctly, you may need to change the data type.

Here are a few other things to keep in mind as you begin working with text values.

- Uppercase letters are not the same as lower case letters to a computer. This is important for matching values and for finding or replacing values in a dataset.
- Numeric characters that are in a string data type are treated as strings, not numbers. This means you cannot perform mathematic calculations on these values without changing the data type.
- Typos, slang terms, and homophones can have a huge impact on text. When working with text data, remember that it is possible the data needs to be cleaned beforehand.
- Spaces and punctuation count as characters too. This comes into play for character lengths, returning characters based on position in a string, and in finding leading or trailing whitespace values which can affect matching.

Boolean Data Types

Another handy data type for information that can only have two values is Boolean. In addition to being fun to say, this data type uses values of zero and one to represent true or false conditions. Boolean data can be a handy way to flag records that meet a certain condition, like if eye color is green. Fun fact, the power button symbol is a combination of 1 and 0 to illustrate that something is either on or off.

Data Essentials

DateTime Data Types

Marking moments in time can help us to order other datapoints to understand trends that would otherwise be missed. For instance, imagine a dataset that shows the altitude of a helicopter. Is this helicopter going up or coming down? If we know when these measurements were taken, we can order the datapoints to show a trend in the data. It's definitely coming down.

Did you know that the date is recorded differently in different cultures? For instance, some places put the current year followed by the month and then day, while others put the year, day, then month. There are other ways of timekeeping that differ entirely, for instance the Wareki calendar in Japan. One of the primary reasons to record dates is to understand the chronological order of data, so we use standardized data types to help.

The most commonly used data type utilizes a standardized method of displaying the date as four-digit year, hyphen, two-digit month, hyphen, two-digit day. For added specificity, we may also mark the time of day by including a value for time in a two-digit hour, colon, two-digit minute, colon, two-digit second format. Some other data types will save space by only recording the date or time.

Specialized Data Types

Across various platforms and programs there are lots of different data types in use. There are some which specialize in holding geospatial data for recording locations. Others reference currency and are intended for use in calculating monetary transactions. Blobs, or binary large objects, are typically used to store multimedia files, like images or audio.

There are many specialized data types, but the key thing to keep in mind is that data types are specific to the type of data you are working with and will vary based on the environment you use. If you run into unexpected errors when working with data, or need to optimize storage or functionality, data types are a great place to start.

Metadata

Metadata

Metadata refers to data about data, which is very useful when you are investigating a new dataset or troubleshooting an issue. While there aren't strict rules about what constitutes metadata, it may be easiest to think of metadata as any attribute or characteristic of data. Examples include: the name, size, and location of a file, which user created the file, the default program for interacting with the file, etc. But metadata isn't limited to files, it also applies to the data values within a file. For instance, the data type or character limit of a given column or value can be considered metadata.

These pieces of data may not seem very useful on the surface, but they can be invaluable when things don't work as expected. Whether you are manipulating data or merging files,

understanding the structure of the files is important. Metadata can help prevent you from making incorrect assumptions about the data. It is also helpful when trying to evaluate a dataset from an external source, since you didn't set up the data.

For instance, imagine you trying to sort a list of numbers in increasing order, but the results are not correct. What could be the issue? Investigating the metadata may reveal that the data values you read as numeric were actually strings!

Visualizing Data

Visualizing Data

You're probably familiar with the saying "a picture is worth a thousand words." When it comes to data, there is merit to the saying but it depends on the image, and your ability to interpret it. Converting a dataset into an image is referred to as visualizing a dataset. It's a great way to quickly understand the scope of a dataset, find trends, and detect unexpected values, known as outliers. However, not all forms of visualization are equally suited for a given dataset. We use different types of charts and graphs to highlight different aspects of a given dataset. Let's take a look at some commonly used charts.

Bar charts are very common and typically used to display summary information of categories within a given dataset. The chart consists of bars placed over two axes, with one axis representing the various categories and the other representing the numeric values. Each bar represents a category in the data and its length indicates the numeric value associated with that category. Data for a bar chart usually consist of one qualitative column and another numeric column. Sometimes a second column of qualitative data can be used to group bars for easy comparison.

Line charts use the values from two columns of data to plot coordinates on a graph. The name line chart is used because the coordinates, or dots, are connected by lines after being plotted. These charts are often used to show trends in a numeric column of data over time, but they can also be used to show relationships between two numeric columns of data.

Scatter charts, also known as scatterplots, are used to display relationships between two numeric columns of information. The chart is made by creating two axes, one for each of the numeric columns, and then points are plotted by using the values found together on a row in the dataset. These are an easy way to investigate relationships because they are visually revealed through patterns in the chart.

Area charts provide a unique view the relationship between categories of data by stacking line charts and filling the area under them. These charts are created using two axes with one representing categorical or datetime information and the other representing numeric values. While there are several different types of area charts, typically they are used to provide perspective and compare values between two or more categories on a single graphic.

Data Essentials

Box & Whisker plots are a way of showing the distribution of values for a given column. This is a good way to see the range of values in a column and understand how spread out the values are from each other. A box & whisker plot is constructed using 5 values that describe the data. It's easiest to understand these values after sorting the data in ascending order. The first two values used are the maximum (or highest value) and minimum (the lowest value). We also use the median, the middle value of the dataset, which splits the data into two halves. The median value of each of the halves are then calculated and added to the plot. The 1st quartile value is the middle value of the lower half of the data set, and the 3rd quartile is the middle of the upper half of the data. A box is created by making a rectangle out of the two quartile marks and whiskers are connected from the quartile markers out to the maximum and minimum values.

Pie charts show how much of the total (what percentage) any given category makes up. These charts are circular, and use lines drawn from the center to slice up the pie into segments that represent the share of each category. They are similar to a bar chart but offer a different perspective, illustrating the relative size of categories within the entire dataset.

Heatmaps take many forms but generally are used to show relationships by using a color spectrum (often blue to red with white showing neutral). These charts are often used to show geographic information (like area occupied on a sports field) but they can also be used to show relationships between different columns of information.

Hierarchy diagrams are used to display relationships that have different levels, such as a company's organizational structure. Often one element branches into several other elements, which have their own branches. This type of information can be difficult to reproduce in structured tables as each level of the structure requires its own column and the information at higher levels will be repeated for all rows. They can also be awkward for data that fits into multiple categories.

Now that we've taken a look at some common graphs and charts used to visualize data, here are a few things to keep in mind going forward.

Compound Graphs

First, there are charts called compound graphs that are used to put more information on display in a single graph. These graphs will often use colors to represent different groupings of information or multiple charts can be stacked on top of each other. Sometimes they use multiple y-axes to show values with different units or ranges. These graphs can be dense and take a little time to understand so like any graphic, don't rush through them.

Look at the Data

Second, don't make assumptions about graphs you didn't create. Graphs are an intuitive way to share data and that can lead people to make assumptions when interpreting the image. Ask

questions about the units of measure, for instance "Is this in pounds or kilograms?". Also ensure the axes are continuous as many graphs will skip over a broad range of values to only show the numeric ranges present, which can be disorienting.

Take a moment to review the graph on the right. Do you notice anything interesting about it? It can be easy to miss but the scale of the bars in the graph are off. For instance, the 2021 points value (275) is more than 9 times the value from 2017 (30) but the graphic underrepresented the difference by showing a bar that is closer to 4 times the size of the bar for 2017. You can still understand the trend, but the visualization is a little misleading.

Dashboards

Lastly, remember that graphs are simply a visual representation of data, and you can always ask to see the data for yourself. There are many bad graphs created, intentionally or accidently, so look at the data and decide for yourself! Good visualizations tell the data's story without need for explanation and that requires picking the right type of graph for the data and supplementing it with additional information.

Visualizing data is primarily done to share the insights with people who either don't have access to the raw data or just want to understand the information contained in the data. Often, these visualizations are collected and sent out in batches via reports or emails. But some cases require quicker updates to the visualizations based on fresh data or require interactivity to allow the viewer to get more data. In those cases, a dashboard can be helpful. Dashboards are usually web based, or live on a local network, and provide a view of the most important visualizations at a glance. Sometimes these are displayed in public areas, or they may be self-service on a website or app.