Data Analysis

A simple definition of data analysis is that it is the process of sifting through data to find useful information.

Depending on the situation, the “sifting” of the data can become a complex operation. It may involve inspecting the data, cleaning the data, transforming the data, or other operations that make finding useful information easier.

Data analysis is widely used in business, industry, research, and education. In recent years the process of collecting and analyzing data has become an industry of its own.

Vocabulary

- Central tendencies (measure of) - Central tendencies are the typical statistical measures of mean, median, mode, standard deviation, variance, and interquartile range (IQR).
- Dashboard - A dashboard is a visual representation of data and data analysis using related charts and graphs. A dashboard can be static or dynamic. Interactive dashboards as the most common form.
- Data analysis - The process of sifting through data to find useful information.
- Data cleaning - Data cleaning is the process of correcting errors and compensating for missing data.
- ETL (Extract, Transform, Load) - ETL is the process of extracting data from a source (usually a file or stream of data), transforming it so that it suits a specific purpose, and loading it into a database or system for analysis.
- Jupyter Notebook - Jupyter Notebook is a popular environment for running Python code as the code and the results are contained in one document. Jupyter Notebook can be used for labs and studies that require repeatability and independent confirmation of results.
- KPI - Key Performance Indicators, or KPIs, are data points that are essential to analyzing the dataset. For example, a dataset covering student performance may use grades and attendance as KPIs. A business may use sales, profits, and inventory.
- Python - Python is a general-purpose computer programming language that uses libraries such as Numpy, Pandas, and Seaborn to perform data analysis and visualization. It can be used as a procedural language or as a scripting language. There are several IDEs for Python as well as command-line versions.

- R - R is a computer programming language specifically designed for processing statistical information. R is popular among non-programmers for its built-in statistical processing capabilities as well as being a scripting language where one line of code can perform complex analysis. R Studio is a popular IDE for R.
- Spread - The spread of the data includes IQR, identification of outliers, and sigma values (standard deviations from the mean). Outliers can exert considerable influence, or leverage, on a dataset. Eliminating outliers is an important part of data analysis.

The Steps in Detail

1. **Specifying Requirements** (setting the goals)

Each project has one or more goals. A question to answer; a hypothesis to (dis)prove; an unknown dataset that needs to be analyzed. Knowing what you need to end up with helps you determine what you need to do to get there.

1.1. Know your audience and your goals

For any data visualization project, you must determine what your goal is. Knowing your audience is key. Take data from an assembly line. The data includes hours online, time and date stamps for periodic samples, error rate in samples, machine ID, operator ID, etc. What data would you include for company leaders? How would it differ if your audience were the assembly line maintenance technicians?

No matter what your audience, you want to keep your analysis to the facts found in the data. Unless specifically requested, you should not include forecasts or what-if scenarios. It may also be tempting to combine data collected from multiple sources, but multiple sources will rarely have the same data collection techniques. However, it is possible to distill a common data set from multiple sources through the data transformation process.

2. **Collect Data**

Data collection (see the module on this topic for more information. Data collection is where a data visualization project often begins. You must collect data and transform it into something useful. For example, a CSV file with weather data would have to be stored into a data structure where it could be easily manipulated. Hand-collected data must be cleaned (checked for errors) and entered into digital format. Fortunately, there are also public resources that have data especially collected for use in data visualization. (See the resources list.)

Just the facts! Resist the urge to editorialize or embellish data, especially if the results seem at odds with “conventional wisdom”. For years temperatures of patients were recorded as 98.6 degrees Fahrenheit regardless of actual readings (unless they were ill) because that was the conventional wisdom about “normal” temperatures. However, studies showed that the average temperature was about a degree cooler and varied with age, time of day, etc. Let the data guide you and not the other way around.

3. Data cleaning and ETL

Data cleaning can take up a large amount of project time. Many datasets are old, incomplete, or contain contradictory data. The process of cleaning data is correcting for such deficiencies and providing a complete dataset with no missing values.

Accounting for missing data is an important step in the data cleaning process. The key is to clean the data in such a way that you do not change its essential nature or its statistical profile. For example, replacing missing values can be done by omitting records with missing fields; replacing missing data with average values; replacing missing data with null values; etc. Each choice has its own benefits and drawbacks.

See the module on Data Cleaning for more information on that process.

Data transformation is also a powerful tool for data analysis. Transforming data is “reshaping” it to meet your needs. In the ETL process, the “T” is for transformation. Sometimes the transformation is necessary: you’ve uploaded CSV data and need to store it as structured data, for example. Other times you are transforming data so that it is in a particular format, such as separating a full name into first and last names.

Data transformation can also be done on the fly. For example, if you are checking a live stream of posts to a social media site, you can filter them for keywords, names, etc. and/or convert them to a different format, such as structured data.

4. Types of analysis.

There is some debate on how many types there are and what to call them!) These forms can be adapted/combined as needed for a particular project. Think of them as tools in your analysis toolbox. Each one has a specific purpose, but you can combine them to solve more complex problems. See the “Types of Analysis” module for more information. Common types of analysis:

- Descriptive Analysis–examining datasets and basic statistical information
- Diagnostic (Inferential) Analysis–finding the “why”
- Predictive Analysis–trends and future planning
- Prescriptive Analysis–optimizing choices (actions)

- Exploratory Analysis–discover relationships between variables

5. **Displaying the data and the analysis.**

Interpretation and publication of data analysis (reports and dashboards) is the final step in data analysis. Summary statistics and visual displays of data and analysis allow end-users to benefit from the data analysis that has been performed. How you present the data and analysis is important. You must provide data and analysis in a form that allows end users to create actionable items that follow up on your analysis.

There are many tools to assist with data analysis and visualization including MS Excel, R, Python, Tableau, KNIME, etc. Most are either free, have free versions for education, or low-cost editions for home & family use. Your educational institution may also have access to SAS, SPSS, Dataiku, or other professional packages.

6. **Drill-down versus scrolling**

When presenting your data and findings, you should make a choice whether to ‘drill-down’ into the data or scroll to see more data/details. There are times when there is a clear choice, such as using scrolling for time series data and drill-down for geographic data.

How to use your dashboards and reports should be obvious to the user(s). So controls should be obvious and instructions embedded in the chart information: “scroll to see more” or “click on a state to see more”.

The goal should also be limited to 3-5 items per screen. You will have to decide whether your dashboards are “drill down” or “scrolling” to show more information. There are applications for both, such as using drill down for Census data (national, state, local) and scrolling for time series data, such as historical oil prices. You can also have combinations, such as a scrolling window within a drill-down dashboard. Examples of displaying data and analysis:

- Charts and Graphs

  A key part of Data Analysis is a visual representation of the data and your analysis. Anscombe’s Quartet is a famous example of data that appears to have the same statistical profile, but when visualized shows very different statistical profiles. Just looking at the *numbers* gives you an incomplete picture. Choosing the right visuals is important for conveying an accurate representation of the data. Some charts are better than others for specific topics. For example, measures of magnitude (length, totals, area, etc.) are best shown with bar charts. Correlations show best with scatter plots. Spread shows well with box-and-whisker charts.

- Dashboards

Dashboards are popular for showing results on a single page. They are frequently used for high-level overviews of data and feature drill-down or scrolling to get more detailed information.

- In-line with Code (Jupyter Notebook, R Studio)

  Jupyter Notebook facilitates a combination of code and graphics. It is easy to create lab report style publications with Jupyter Notebook and the inclusion of the code used to generate them allows people to recreate the Data Analysis process.

  R also allows you to embed charts and graphs with the R Studio IDE.

Data Analysis as a Class Topic

Data Analysis can be done as a module in a larger class or as an entire class unto itself. Aside from the information and processes above, some of the topics include:

1. Anscombe's Quartet

   Anscombe's Quartet is a famous lesson that demonstrates why visual analysis of data is crucial. Four sets of data appear to have the same statistics for center/spread. However, once they are graphed dramatic differences can be seen. This lesson emphasizes the importance of visual analysis.

2. Business KPIs

   KPIs (Key Performance Indicators) are often used in business. They are essential parts of data that can show at a glance whether projects are on time/budget/etc.

3. Trend analysis

   Trend analysis is a great way to show change over time. Great for "subway" charts (grouped line charts).

4. US Census

   The US Census is a treasure trove of data. Use it to demonstrate a specific data analysis method or a combination of them.

5. Explaining the basic concepts of Python and R and using Python and R for Data Analysis

   Popular programming languages frequently used in data analysis include R and Python. Both are easy to use and ban be used interactively or to create stand-alone programs.

6. Identifying the right tools, concepts and functions that are required for Data Analysis

With a field as large as Data Analysis, there will be a number of books and tutorials available. Try a few of the more popular ones (especially the free ones) to see what is the best fit for your classes. Note that several of the commercial products have free licenses for educational use (Tableau, PyCharm, etc.)

Classroom Project Ideas

1. Generate Anscombe's Quartet from a CSV file. Discuss how analysis differs once you see the data in a chart.
2. What might be some "business" KPIs for the class? Do only grades matter? Would a correlation analysis find other potential KPIs?
3. Have students perform trend analysis on their own grades. What grade is predicted for the end of the term? Can a trend change? (A great project to discuss outliers, too.)
4. Analyze the Iris data set. Start with a simple scatter plot of length versus width, then have students pick another chart type or two to perform further analysis. What did they find out? Follow-up: pair (correlation) plots using Iris data. Is it easier to spot correlations of variables with a pair plot?
5. Have students explore the US Census website. What data visualizations are the most interesting? Why? Have students pick one visualization and redo it with a different chart type. (You may want to choose the chart type for them.) Is the new chart better or worse? Why?
6. Create a science project (growing plants from seeds, etc.) and collect data from it. Create a dashboard to display the information. This is a great project to explore creative ways of charting, including using plants as data markers and plant-specific colors as a theme.
7. Have students collect data about something they do every day or week, such as brushing their teeth or doing yard work. Students should create a template for data entry that contains at least five observations. (Day, time of day, duration, weather if outside, etc.) Compare results weekly and for the entire term using dashboards.
8. Create a word cloud from a book they are reading for (another) class. The book should be available in a searchable format to make this easier. (PDF is popular and you can use web scraping for HTML formats.) This is a great assignment for out-of-the-box thinking. Word clouds are an obvious choice, but what might other charts reveal? What are the limitations of word cloud charts?
9. Statistics Education Web lesson plans: https://www.amstat.org/education/stew/statistics-education-web-(stew)

Resources

1. Free data viz tools. https://rigorousthemes.com/blog/best-free-data-visualization-tools/
2. Natural language processing with Python and the Natural Language Toolkit. https://www.nltk.org/book/
3. Extracting text from PDF file tutorial (Python). https://www.geeksforgeeks.org/extract-text-from-pdf-file-using-python/
4. UCI Machine Learning website (beta). https://archive-beta.ics.uci.edu/
5. Data + Science data visualization website. https://www.dataplusscience.com/insights.html
6. Makeover Monday. https://data.world/makeovermonday
7. Choosing the right chart for your data. https://help.tableau.com/current/pro/desktop/en-us/what_chart_example.htm
8. Big Ideas. https://www.youcubed.org/data-big-ideas/
9. American Statistical Association: https://www.amstat.org/education/k-12-statistics-education-resources-
10. Census Bureau data: https://www.census.gov/data/datasets.html
11. Data Science with Python tutorials: https://www.w3schools.com/datascience/ds_python.asp
12. R Studio: https://www.rstudio.com/
13. Anaconda Python distribution: https://www.anaconda.com/products/distribution
14. Data Science glossary: https://datascienceglossary.org/

This material is based upon work supported by the National Science Foundation under Grant DUE 2055411.Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.