Data Analysis Fundamentals

In this module, we will discuss storytelling and communication: how effective data scientists explain and interpret their results, as well as how to communicate findings accurately to stakeholders to inform business decisions.

Topics covered

- Data Analytics Life Cycle
- Meeting stakeholder needs
- Communicating Data Science results
- Storytelling and communication
- Creating effective visualizations

Skills learned

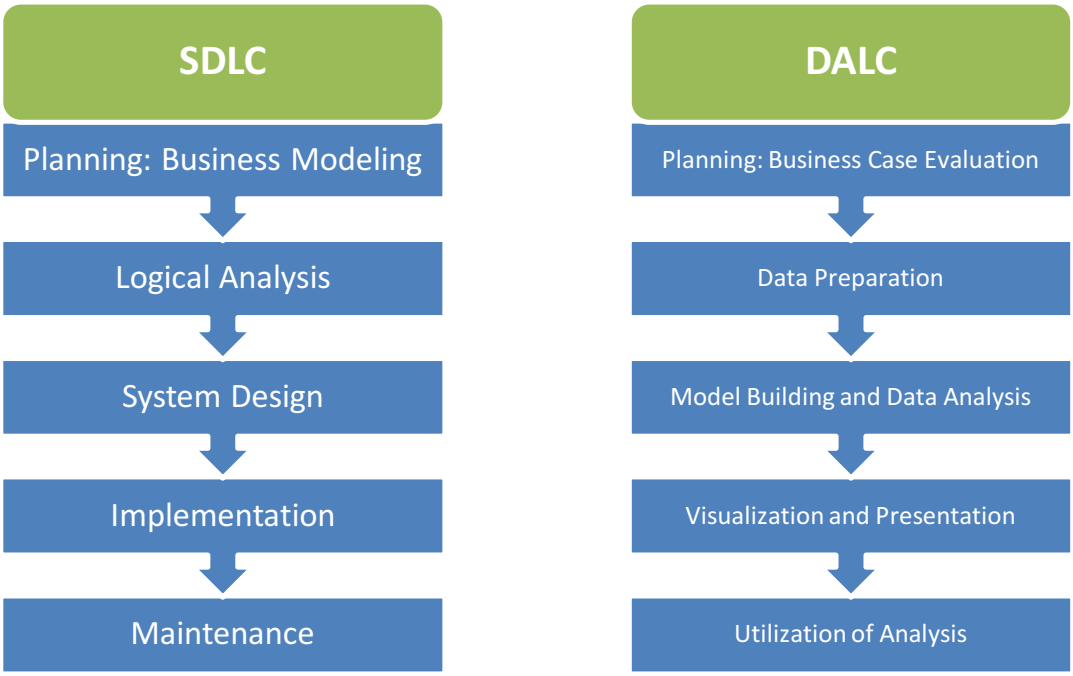
We will learn about the Data Analytics Life Cycle. We will discuss the techniques that are needed to join a data science team and to contribute quickly and effectively to the team. After learning about the different kinds of data that characterize big data, we will delve into approaches to visualize and present that data effectively to the various stakeholders.

Data Analytics Life Cycle (DALC)

The goal of this module is to get you up to speed quickly. You might be part of a larger data analytics team or you might be starting your own team of one. No matter the situation you face, our goal is to have you hit the ground running. In order to do so effectively, we will start by giving an overview of what's involved in a data analytics project.

The intended audience is anyone that can be in a data analytics team, including business analysts, data analysts, data scientists, machine learning experts, business intelligence experts, etc. It can include members of the technical group, administrative staff, and management team. Regardless of your particular role, the stronger your technical background, especially in programming and probability and statistics, the better positioned you will be to make significant contributions.

The Data Analytics Life Cycle (DALC) is analogous to, but not as well-established as, the Software/System Development Life Cycle (SDLC). In a very real sense, any such life cycles can be seen as a customization of the traditional scientific approach for a particular problem domain, like software or big data. The traditional SDLC and the DALC can both be seen in the figure below.

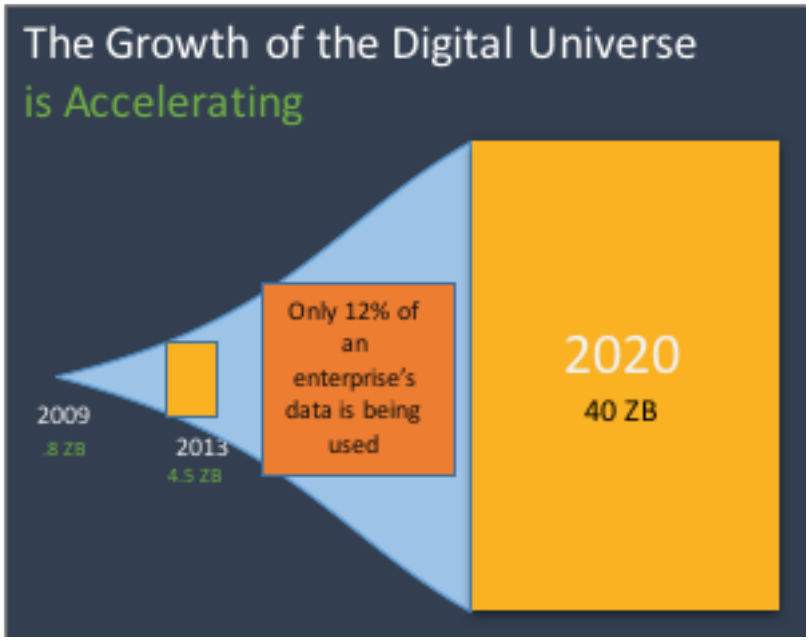


SDLC (left) vs DALC (right)

It's important to note that the DALC has not been formally established yet; although there are a number of suggested DALC suggestions[1], the above abstraction is a good representation of the steps involved in any data analytic task. The first step, Planning, is arguably the most important one as this encapsulates learning about the business domain, framing the business problem, and formulating the initial hypotheses. Although these steps are represented sequentially, the DALC, like the SDLC, is not purely sequential and any and all steps can be iteratively revisited in preparing the final analysis; in fact, it is quite common to move from any one phase to another phase and back again. The overall approach, as shown in the DALC above, represents a best practices approach for an end-to-end analysis of any project involving big data.

Dealing with big data is increasingly necessary as the growth in big data over the last few years is nothing short of remarkable; the projected growth over the next few years is even more spectacular. The following figure from the IDC Digital Universe Study from 2013[2] projects that the amount of data will increase 10-fold from 4.4 trillion GB to 44 trillion GB. Although the scale, or size, of data is one aspect of big data, other characteristics like the volume of data (how fast it's increasing) and the variety of data (the different formats and media of the data) are also cause for concern.

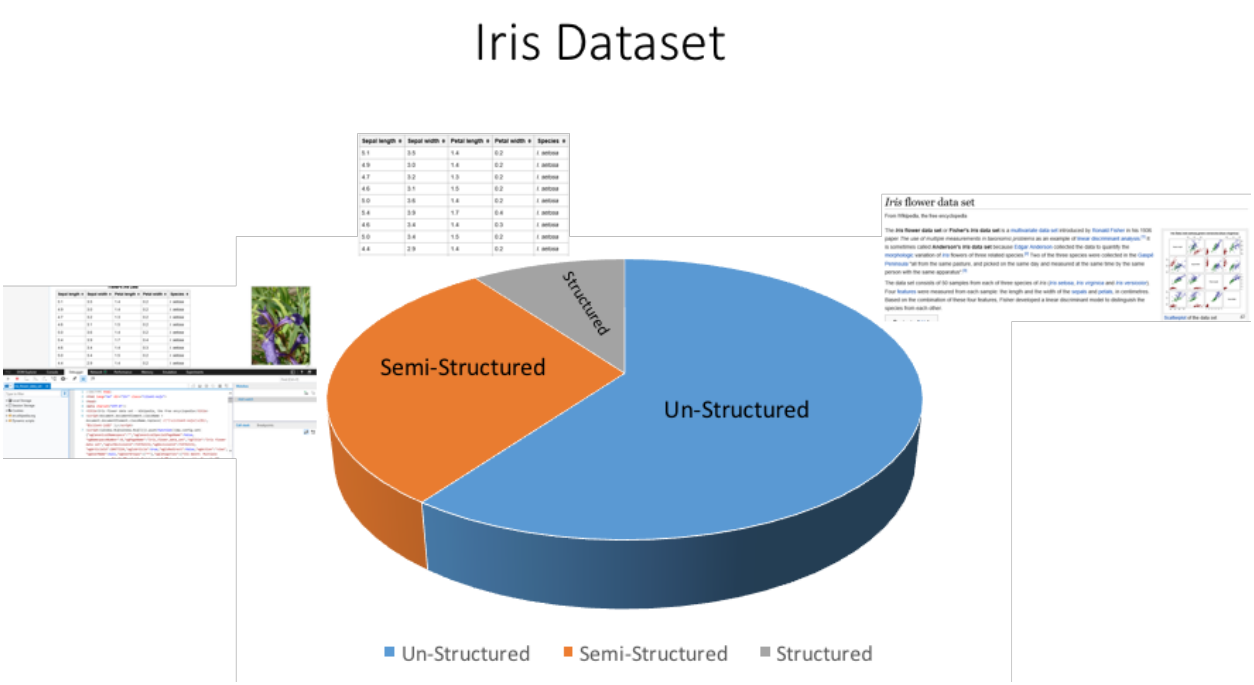
[1] Source: http://www.informit.com/articles/article.aspx?p=2473128&seqNum=11
[2] Source: http://www.infodocket.com/2014/04/16/how-large-is-the-digital-universe-how-fast-is-it-growing-2014-emc-digital-universe-study-now-available/



IDC Digital Universe Study from 2013 – Figure recreated from http://image.slidesharecdn.com/webinar-fastandfurious-frompoctoenterprisebigdatastack20140424-140425115111-phpapp01/95/fast-and-furious-from-poc-to-an-enterprise-big-data-stack-in-2014-6-638.jpg?cb=1398436798!

Structure of Data (DALC: Stages 1 & 2)

Data comes in many forms: it can be highly **structured**, like typical database tables or excel spreadsheets; it can be **un-structured**, like emails, documents, images, etc. – anything that doesn't neatly fit into a database; or it can be **semi-structured**, living somewhere between the two, like XML or JSON files. It's important to note that un-structured data, in this sense, does not mean the data itself doesn't have some internal organization; instead, data is called un-structured in the sense that normal data mining tools cannot easily parse it. Most data generated today is either semi-structured or un-structured, as shown in the graphic below.



Structured, Semi-Structured, and Un-Structured Data using the Iris Dataset Wikipedia Page – **Make Bigger**?

In fact, an article written by Seth Grimes[3] suggests that up to 80% of business-relevant information comes in the form of un-structured or semi-structured data. Why is this relevant for us? The type of data we're trying to analyze will determine the kinds of visualizations we can use. Whenever we deal with different kinds of data, we need to ask questions like:

- What kinds of questions can we ask using this kind of data?
- What types of analytics can be carried out on this data?
- What data sources are available?
- What types of data will we have (structured, semi-structured, un-structured)?
- Who uses or consumes this type of data?

Identifying Data Analytics Roles (DALC: Stage 1)

In order to identify who can benefit from an analysis of big data, we need to identify the users and stakeholders. These stakeholders fall under the category of *Business Users*, *Business Executives*, and *Analysts*, either data analysts or data scientists. In addition, any data analytic project team should be composed of the right mix of domain experts, customers, and data analysts, who often serve as the intermediary between the data scientists and project management team. The questions to ask at this stage are:

- Who will benefit from the analysis of this big data?
- Who can conduct the analysis and visualization?
- Who could provide additional insight as consultants?

[3] Source: https://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/

- Who has final authority on the project?
- Who will determine the success or viability of the results and recommendations?

Identifying Stakeholder Needs (DALC: Stage 1)

Once the stakeholders have been established, the next step is to identify their business needs and address them for each group individually. Stakeholder needs are commonly identified in the first step of the DALC, Planning, in which we frame the problem by stating the analytics problem we want to solve and clearly articulating the current situation or state. At this stage, we will also need to clearly state why the problem is important and which of the stakeholders will benefit from its solution. This is usually done by interviewing the stakeholders to determine their business needs. The relevant questions at this stage are deciding:

- What are the business problems we need to address?
- What is the goal or desired outcome of this analysis or visualization?
- Are there any temporal constraints for this project?
- What are the key risks involved in this project and analysis?
- What are the criteria or metrics for success?

Business needs often fall into the following categories:

Optimizing business operations like profitability, sales, efficiency, etc.
Identifying business risks like fraud, churn, etc.
Predicting new opportunities like new prospects, cross-selling, etc.

Determining roles in a data analytic project: One way to determine roles and responsibilities is to build a RACI matrix, which identifies the role categories for a project as indicated below and establishes clear agreement and protocols:

Responsibility: people in this category are expected to actively complete assigned tasks
Accountability: people in this category have ownership of a given task
Consulting: people in this category are domain experts that can be consulted during the analysis
Informing: people in this category are notified once a decision is made

---

Creating Effective Visualizations (DALC: Stage 4)

Although the Planning stage (Stage 1 of the DALC) can be visited multiple times, once the initial planning is done, data sources are amalgamated and data is prepared (Stage 2 of the DALC). At this point, the Data Analysts and Data Scientists engage in Model Building and Data Analysis (Stage 3 of the DALC). Concurrent with the data analysis, we can start to visualize both the data and the results (Stage 4 of the DALC). We can also begin creating presentations for the various stakeholders and decision makers and also iteratively revisit Stage 3 to deepen the analysis, validate our data models, and begin creation of the final reports.

Visualization is inherently dependent on the structure of the data you have to analyze. As was discussed earlier, there are three kinds of data: structured, semi-structured, and un-structured. Each of these different types of data calls for a different approach for presentation and analysis, including the underlying tools.

Visualization contributes to both *Business Intelligence (BI)* and *Predictive Analytics (PA)* and is essential to *Decision Making*, as well. We can differentiate between visualization methods for Business Intelligence and Predictive Analytics by looking at the kinds of questions they ask, the kinds of data they analyze, and the typical analytical and visualization techniques they employ.

Visualization for Business Intelligence elaborates upon questions like: What happened over the last two quarters? How many units were sold in a particular region? Which customers purchased the most units over the holiday season? It usually deals with structured data from traditional sources using manageable datasets that are not inordinately large. Some typical visualization techniques for Business Intelligence include standard reports, dashboards, queries, etc.

Visualization for Predictive Analytics, on the other hand, deals with questions like: What will happen if the current trends continue? What is the optimal scenario for sales in the upcoming holiday season? What causal relationships are there between sales trends and business conditions? It usually deals with structured, un-structured, and semi-structured data from variegated sources that often contain extremely large datasets. Some typical visualization techniques for Predictive Analytics include optimization graphs, forecasting charts, statistical analyses, etc.

In addition, the visualizations generated in Stage 4 of the DALC can feed back into Stage 3 of the DALC (Model Building and Data Analysis) and help answer questions like:

- Does the model output make sense?
- Were we able to validate the model on the test and validation datasets?
- Do we need more data or more inputs?
- Do we need to adjust the parameter values of the model?
- Do we need to modify or replace the model itself?

The results of these visualizations, along with the actionable recommendations, are often packaged into presentations so that the appropriate stakeholder can utilize them for Decision Making. Thus it is especially important to ensure your presentations are effective and impactful.

Preparing Impactful Presentations

Each presentation should be tailored for a specific target audience and a specific purpose. Presentations can be aimed at Business Executives, Business Users, or Analysts, where the amount of business impact (as measured by metrics like risks or ROI) is replaced by technical content (discussing model details and production environment issues), in essence shifting from the "what" to the "how". All of these formats for the presentation should, however, follow a uniform format[4]:

- Start with the answer first.
- Group and summarize your supporting arguments.
- Logically order your supporting ideas.

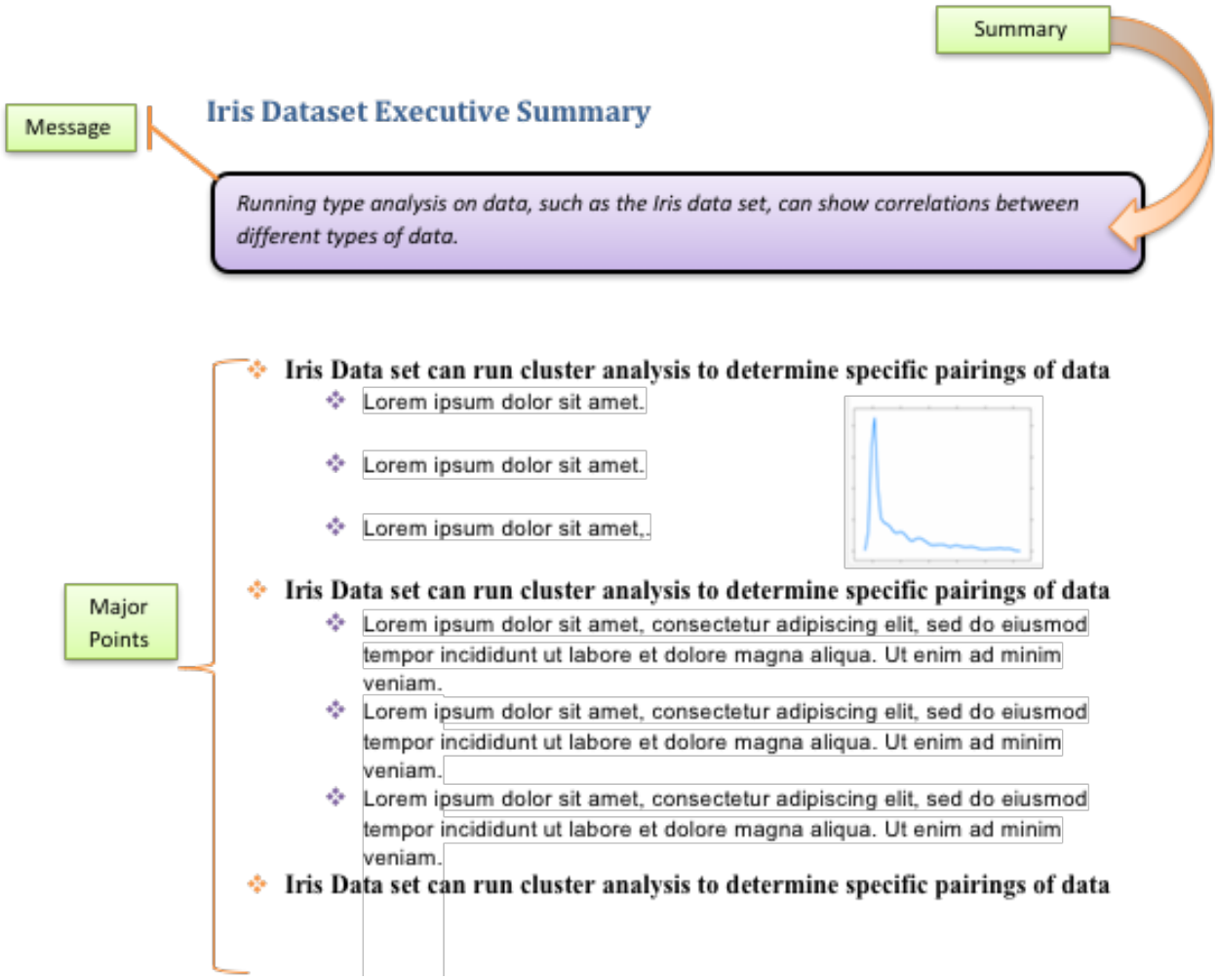
Following Barbara Minto's principles, when grouping and summarizing your supporting arguments, follow the principle of three's: have three sections with each section having three supporting arguments. If your presentations have a temporal component, these subsections should be organized temporally; otherwise, they should be presented in order of importance or, if there's a natural structure, following that structure.

The final report should thus have five sections in all:

[4] Source: The Pyramid Principle from Barbara Minto and https://medium.com/lessons-from-mckinsey/the-pyramid-principle-f0885dd3c5c7#.ubatffkxr

1. Project Goals
2. Main Findings
3. Model Description
4. Findings and Visualizations
5. Actionable Recommendations

The Main Findings section presents the Executive Summary and should be the most concise portion of the presentation and should be a single slide at most, as shown in the figure below.



Sample Main Findings Executive Summary – **Finish customiztion this for Iris**!

The main findings section is the crux of the presentation and the one section everyone will read; in fact, it might be the *only* section some people read. This section should thus clearly and succinctly communicate the key insights and outcomes. The outcomes should be framed in terms of the business value, as shown in the figure above. Following Minto's principles, the main answer is given at the very top in a highlighted box. This answer should, if possible, quantify the results; it should make the business impact concrete by quantifying the value of the work, the cost savings, the revenue, the time savings, or some other benefit for the business practices. Also following Minto's pyramid principle, the main message should be supported with three major points; one of these key points should include a visualization that

summarizes or exemplifies the main contributions, as shown above, as visual imagery often creates a visceral connection with the reader and helps them retain the central message.

Presentation Tips (RESTATE!)

- For Business Executives, the more succinct the presentation, the better. Most executives attend many briefings in the course of a day or a week. Ensure your presentation gets to the point quickly and frames the results in terms of value to the sponsor's organization. For example, if you are working with a bank to analyze cases of credit card fraud, highlight the frequency of fraud, the number of cases in the last month or year, and how much cost or revenue impact there could be to the bank (or focus on the reverse: how much more revenue they could gain if they address the fraud problem). This will demonstrate the business impact better than deep dives on the methodology. You will need to include supporting information about analytical methodology and data sources, but generally only as supporting detail or to ensure the audience has confidence in the approach you took to analyze the data.
- For Business Users, try to utilize imagery when possible. People tend to remember mental pictures to demonstrate a point more than long lists of bullets.
- For Analysts, focus more time on the methodology and findings. You can afford to be more expansive in describing the outcomes, methodology and analytical experiment with a peer group, as they will be more interested in the techniques, especially if you developed a new way of processing or analyzing data that can be reused in the future or applied to similar problems.

Stakeholder-specific Visualizations

Visual presentations can be customized for Business Executives, Business Users, or Analysts. For Business Executives, key points should be supported with simple charts and graphics, like bar charts. All visualizations should illustrate data clearly and help the audience understand the value of the insights.

For Analysts, the charts and graphs should be more detailed and present a higher level of technical depth. Appropriate graphs can be ROC curves, histograms, etc. In addition, visualizations for analysts should also highlight the key variables using visuals which highlight the significance of each such variable. In general, the graphics for analysts should have a much greater granularity and level of technical detail, using visualizations like dot density charts, histograms of data distributions, etc.

Finally, for Business Users, an intermediate level of detail can be used in the visual presentation. Key points should be illustrated using charts and other visualizations and also include the model metrics. In general, the approach for users should reflect the same sensibility as for Business Intelligence. For all users, it's a good idea to focus on clean, easy visuals and, especially for Business Executives, to visualize the key messages that will aid their Decision Making process.

Survey of Data Visualization Tools

The explosion of big data across all industries over the last few years has fueled the need for greater analytics. With this growing demand, organizations are beginning to favour ease of use and visualization in Business Intelligence rather than the more traditional, yet harder to use, Business Intelligence tools and databases. This new crop of tools has an innovative user interface that makes visualization easy. These tools are both open source and commercial and can be supplemented with more traditional business intelligence tools.

Open Source Visualization Tools

- R with packages like ggplot, Lattice, etc.
- python with packages like matplotlib, etc.
- Modest Maps: JS, python, etc.

Commercial Visualization Tools

- Tableau
- Qlikview
- HortonWorks

Traditional Analytic Tools

- Programming tools such as R (RStudio) and python (Anaconda)
- MapReduce/Hadoop
- In-database analytics

Using Charts in Presentations

In a later module, we'll talk about how to view quantitative data and the different kinds of data we'll come across as Analysts. For now, let's focus on the basic kinds of data we'll need to visualize for Business Executives and Business Users, as shown below:

If you want to compare this kind of information….	…consider this kind of chart
Components	Pie chart
Item	Bar chart
Time Series	Line chart
Frequency	Line charts, histograms
Correlation	Scatterplot, side-by-side bar charts

Shown above are some basic chart types to guide you in considering that different types of charts are more suited to the situation depending on the data you have and the message you are attempting to portray.

The table is by no means exhaustive; it is illustrative to convey the most basic data representations, which can be combined, embellished, and made more sophisticated depending on the situation and the audience. Consider the message you are trying to communicate, then choose an appropriate visual to support the point. Misusing charts tends to confuse the audience, so be sure to take into account the data type and message when choosing a chart.

Pie charts are designed to show the components, or parts relative to the whole set of things. It is also the most overused chart. If you are going to use a pie chart, use it when showing only 2-3 items in a chart and only for Business Executives. Bar charts and line charts are used much more often, and are very useful for showing comparisons and trends over time. For bar charts, horizontal bar charts allow you to fit the text labels better and provide more horizontal space to fit them next to a chart, even though many people tend to use vertical bar charts. Vertical bar charts tend to work well when the labels are small, such as when showing comparisons over time using years.

For frequency distributions, histograms will show the distribution of data, and are useful for showing information to an audience of Analysts, either Data Analysts or Data Scientists. The data distributions are typically one of the first steps in visualizing data and preparing for the model planning. When doing correlation, scatterplots are useful to compare relationships among variables.

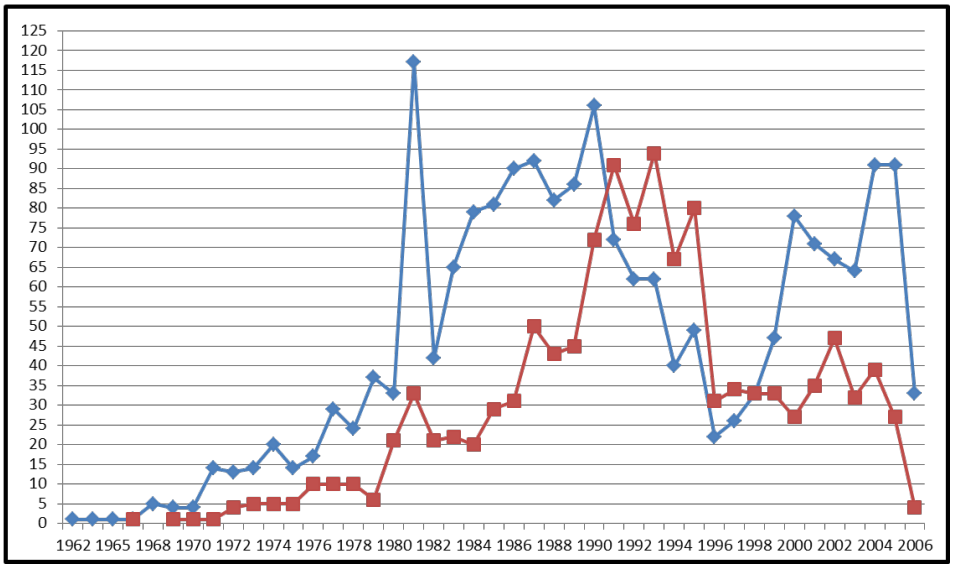
As with any presentation, consider the audience and their level of sophistication when selecting the chart to convey your message. These charts are simple examples, but can easily become more complex with additional data variables, combining charts together, or adding animation where appropriate. When manipulating such charts, one of the best recommendations is to consult Gene Zelazny's classic work, Saying It With Charts, which is also summarized online[5].

Cleaning Up Charts and Visualizations (REDO FOR IRIS)

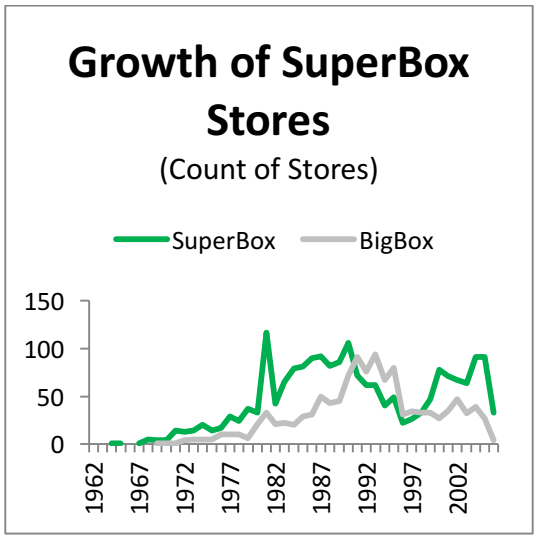
Example 1

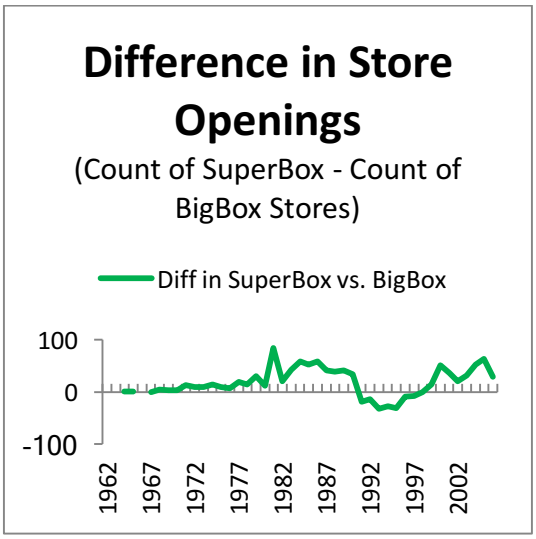
Chart junk

1. Horizontal Grid Lines
2. Chunky data points
3. Overuse of emphasis colors; lines & border
4. No context or labels
5. Crowded axis labels



[5] Source: http://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html

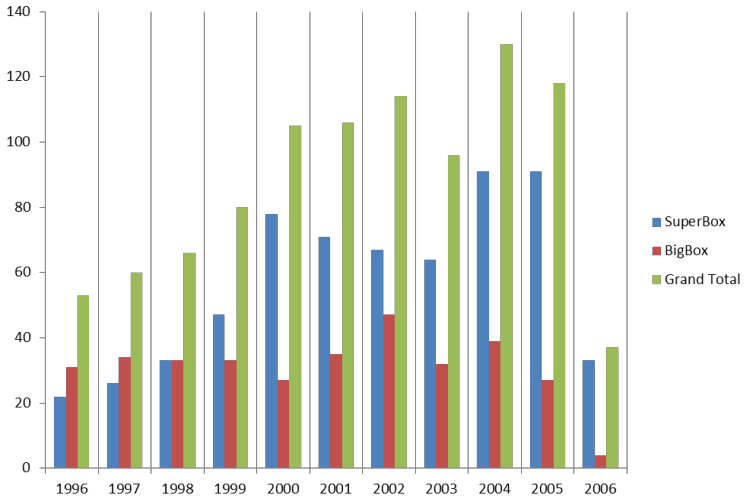


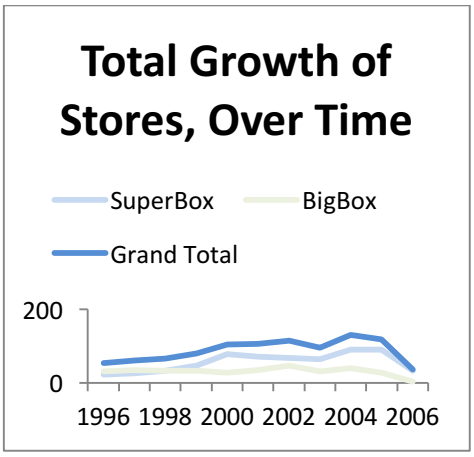


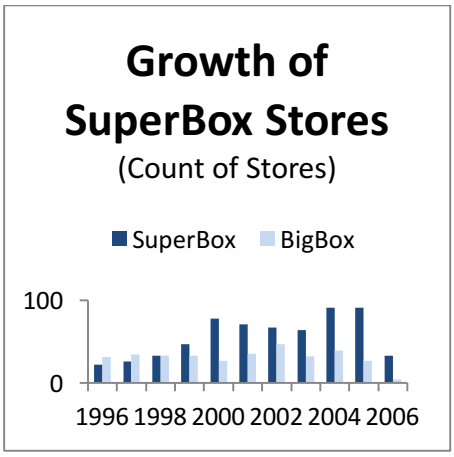
Example 2

Chart junk

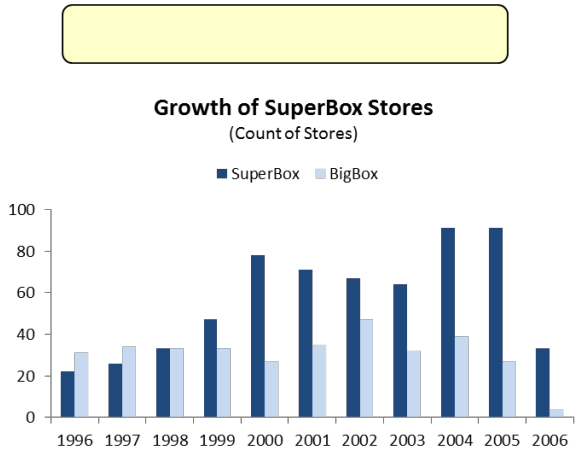
1. Vertical Grid Lines
2. Too much emphasis colors
3. No chart title
4. Legend at right restricts chart space
5. Labels are too small

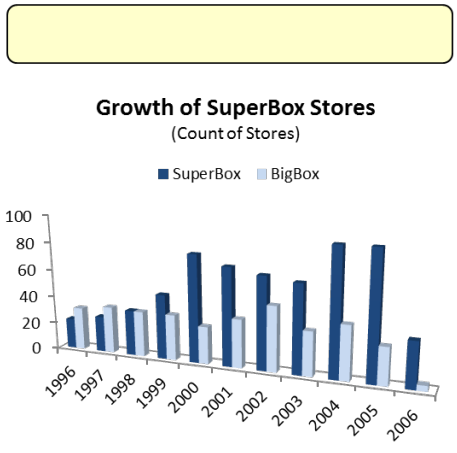






Example 3





Key Points

- Remove distractions
    - Minimize “chart junk”
    - Data-Ink Ratio
- Choose the simplest, clearest visual for the situation
    - Strive to illustrate your points
    - Charts should serve to reinforce your key points
    - Charts vs. Data Art
- Use color deliberately
    - Emphasis Colors vs. Standard Colors
    - In most cases, less is more
    - Focus on the contrast
- Context
    - Consistent scales, labels, axes

    - Using logs vs. raw values to show differences

In the next module, we'll learn about the different kinds of quantitative data, how to visualize it, and start to delve into both R and python to analyze the Iris dataset.

References

Here are some references to deepen your understanding of the best practices for visualizing data:

- Edward Tufte's Envisioning Information: this is a pioneering text on presenting information in graphical form; accessible and comprehensive, it beautifully illustrates the foundational principles of presenting complex material by visual means.
- Stephen Few's Information Dashboard Design: this is written by one of the experts in data visualization; it gives an introduction to the fundamental principles of data visualization and design and is full of examples of good and bad dashboards.
- Barbara Minto's Pyramid Principle: this is a foundational work for constructing logical structures for presentations and weaving a story out of the disparate analytical results. The recommendation is to have three sections for the presentation wherein each section has three main points.
- Gene Zelazny's Say It With Charts: this is a lovely reference book for picking the most appropriate graphical form for your data to ensure information is clearly conveyed.
- Garr Reynolds' Presentation Zen: this has many examples for helping you convey ideas simply and clearly using appropriate imagery.