Is Your Visualization The Best or A Mess?

De'Ambra Unuigbey & Steven Collins
Data, Analytics, & Institutional Research
University of North Texas, Denton TX USA

Abstract

Ever wonder if your visualizations are painting a clear picture or conveying the correct message? Often in a quest to tell a story the meaning gets lost in the details. Charts are poorly designed, too busy, or are just simply the wrong chart. The design can lead to false conclusions as opposed to valid depictions of what's really going on. This paper will provide key insights on how to successfully tell a story with visualizations while avoiding common pitfalls, such as designing for yourself instead of the audience or putting too much data into one chart. We will provide good design practices including using the correct charts and graphs to display data that conveys a clear story without the ambiguities of poor color choices, sizing, titling and many other errors that frequently occur. Overall we will show how to take visualizations that are a mess and make them the best leaving the audience well informed.

Intro

How easy it is to create an anecdotal visualization(s) is a matter of conceptualizing the story that the data is intended to convey. In turn, this message helps determine which graphs best depict the information and shape the overall narrative. Conceptualization is what makes a good visualization standout. This paper examines best practices and tools of visualization, identifies characteristics of a good story, and explains how to avoid common chart mistakes. So gather the data and let's dig in.

Tell a Story

Story telling has been a part of culture since humans first used cave walls to paint narratives thousands of years ago. Visualizations connect people from different backgrounds to a central idea or observation. The cogent flow of a storyline draws in and focuses the audience's attention on important findings. Anecdotal visualization is a means of concisely directing information so that it is clear, succinct, and engaging. In short, stories are used in visualizations in order to enhance the attention of the audience to what is important.

The Audience

An assignment requires analyzing the current fiscal year's position to create a performance forecaster. What is the first step? A good first step is to identify the audience and determine what their individual expectations are - what do they expect to see, what is their daily processes, and where would they like to be next year. Avoid making assumptions about what the audience may want to see, as this leads to self-visualization. The ideal means of directing attention is to design for the audience. Communicating with the audience in advance is essential in determining the best possible means of presenting the data. A story guides the audience through complex sets of data by making the information relatable. Essentially, allow your audience to establish a starting point of common expectations and conclusions, and let your data tell the company's story. This will illuminate misconceptions and steer business in a better direction.

Once all the data has been gathered and there is a good inkling of what the audience wants to see, what is next?

Explore the Data

The next step is to determine the best means of presenting an anecdotal visualization. This involves exploring the data and discovering points of agreement with and departure from the audience's expectations. By honing in on these critical points, the storyline is tailored to the audience. Start by experimenting with different scenarios to determine the storyline(s) hidden in the data. A couple of good tools are Data Exploration in SAS and MS-Excel. SAS Explorer will automatically select the best object for the visualization when data items are dragged to the center work-space. In Figure 1.1, there are four types of visualizations that are great for exploring and determining how best to display the data. Crosstabs and word-clouds provide drill-through capabilities based on a hierarchy. By drilling down, narratives can emerge that provide clarity of how data interacts or breaks-down. Moreover, SAS Data Explorer allows local data files, such as Excel, to be uploaded as well as from servers, Hadoop or applications like Facebook, and Google Analytics.[1] Additionally, Explorer has great analytical tools, such as Sankey Diagrams, Decision Trees, and Linear Regression.

[1] SAS Visual Analytics, 2018. https://www.sas.com/en_us/home.html.

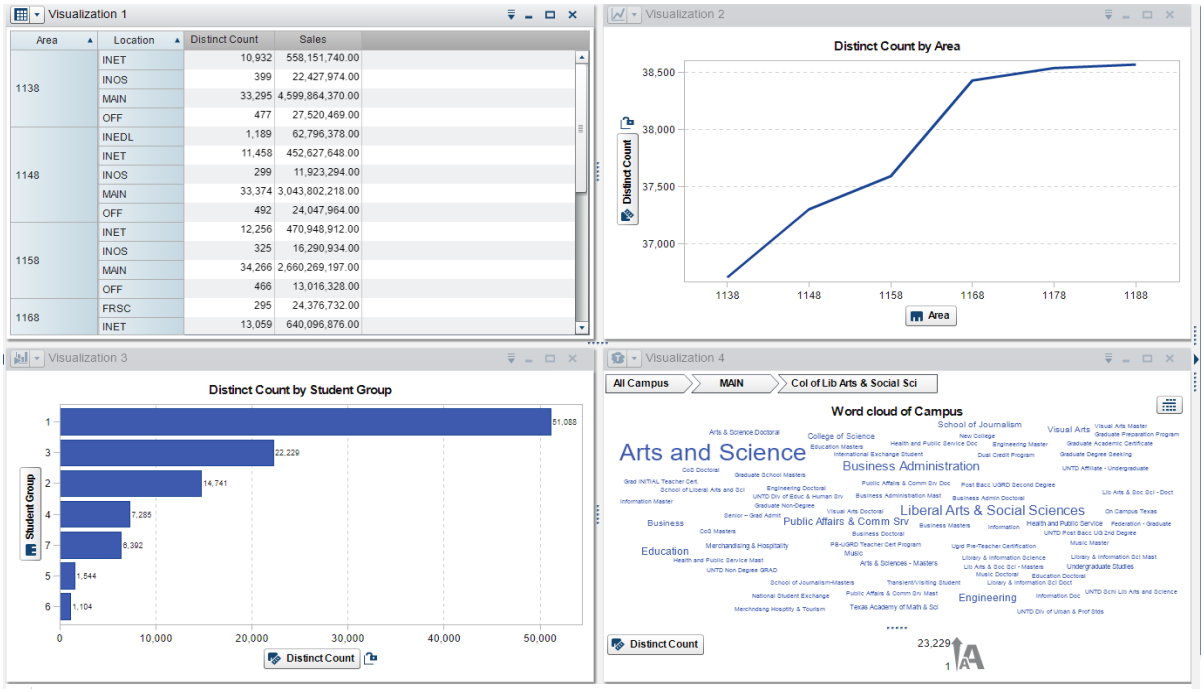

Figure 1.1. SAS Exploration of Fictional Data

In excel, slicers added to graphs can provide the same multi-level analysis as a crosstab or word-cloud (see Figure 1.2 below as an example).

GTP_START_TERM	(All)						
GTP_END_TERM	(All)						
Row Labels	**Sum of SCH_LOW**	**Sum of SCH_UPPER**	**Sum of SCH_MAST**	**Sum of SCH_DOC**	**Sum of SCH_SPEC_PROF**	**Sum of NTSR_CBM_EXC_SF_SCH**	**Sum of Grand Total**
Administrative Group	-	-	-	-	-	-	-
College of Business	3,891	7,749	81	-	-	114	11,835
College of Education	4,090	4,264	37	-	-	100	8,491
College of Engineering	2,553	2,297	18	-	-	96	4,964
College of Information	324	792	-	-	-	18	1,134
College of Mrch, Hosp, Tourism	3,064					28	4,610
College of Music	3,340					19	4,483
College of Visual Arts & Dsgn	5,254					6	6,659
Honors College	139					-	203
Mayborn School of Journalism	1,269					3	2,613
Col of Health & Public Service	2,774					54	6,678
Col of Lib Arts & Social Sci	30,348					162	40,899
College of Science	16,986					245	20,526
New College	9					9	189
College of Arts and Sciences	-					-	-
Grand Total	**74,041**					**854**	**113,284**

Figure 1.2. Excel Pivot Table and Line Chart of Fictional Data

Data selection is important, as well is honesty. Designing for the audience is not a matter of telling the story that the viewers want to see, rather telling a story that conveys clarity and understanding. Avoid embellishments or omissions that skew or otherwise bias the data. Data selection should include information that supports and rejects the beliefs of stakeholders, and presents unexpected discoveries.

Vicissitudes in graphs may indicate important changes, likewise it can signal errors in the data set. Data preparation like scrubbing and normalizing, adjusts for missing or incomplete information, misspellings, and data anomalies. However, missing data can be a clue to an important part of the story. The motto garbage-in garbage-out applies here. Flaws in data collection cause graphs to sew more uncertainty than clarity. Conversely, clean data conveys trustworthiness.

The significance of data governance is often overlooked. Consistency across divisions in definitions, terminology, and data collection circumvents internal misunderstandings that may diminish the value of data analysis. Therefore, data integrity is important as data items could mean something entirely different across departments. For example, the data term "Client" to a sales manager is a customer, yet to human resources supervisor it is a person who is an employee, auditor, or sub-contractor.

Storyboard

Most artists, authors, and presenters capture the full story by outlining the final product through a storyboard. Storyboards are graphical and visual outlines of how a story is structured. Tools as simple as pen and paper can be used to construct storyboards. Adobe Illustrator can produce a comic strip type of storyboard. Storyboards allow the rearranging of ideas and the introduction of new discoveries so that the visualization has a clear beginning and end. Below is a small example of a storyboard from Cole Nussbaumer Knaflic. Knaflic discusses how she turned a pen and paper mock up into a set of meaningful graphs.[2]

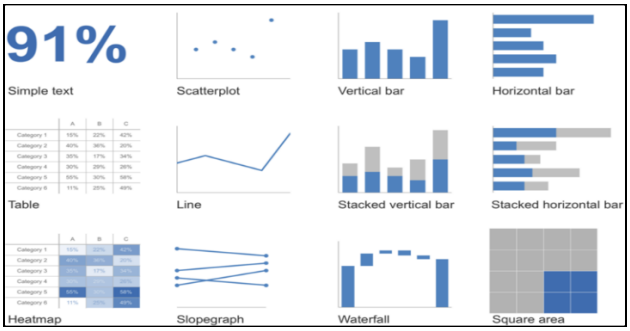

Figure 2.1.[3]

[2] Cole Nussbaumer Knaflic, "How I Storyboard," *Storytelling with Data* (blog), August 25, 2015, http://www.storytellingwithdata.com/blog/2015/8/24/how-i-storyboard.
[3] Cole Nussbaumer Knaflic, *Storytelling with Data: A Data Visualization Guide for Business Professionals.* (Hoboken, New Jersey: John Wiley & Sons, 2015)

Storyboards are useful and can bring data to life. Creating a story brings clarity to visualizations that makes sense to the audience.

Visual Perception

When creating a visualization, be aware of the visual perception. The Gestalt Principles of visual perception are a set of laws arising from 1920s' psychology, describing how humans typically see objects by grouping similar elements, recognizing patterns and simplifying complex images.[4] The idea was that human communication originated from the way they perceived certain structures or items as a whole versus a specific unit within the structure. Figure 3, shows the most common principles: closure, proximity, continuation, and figure and ground. "Closure" is when viewing an incomplete image, the human mind sees the structure as a whole by mentally closing in the gap(s) that is missing. "Proximity" allows us to cluster items that are close together and "Continuation" permits continual follow through of an object the author may not have otherwise filled in. Through "Similarity," the mind assembles portions, which look similar to each other, while "Figure and Ground" allows seeing lighter imageries that are in the background of a darker figure that may or may not have been the original intent of the true image.[5]

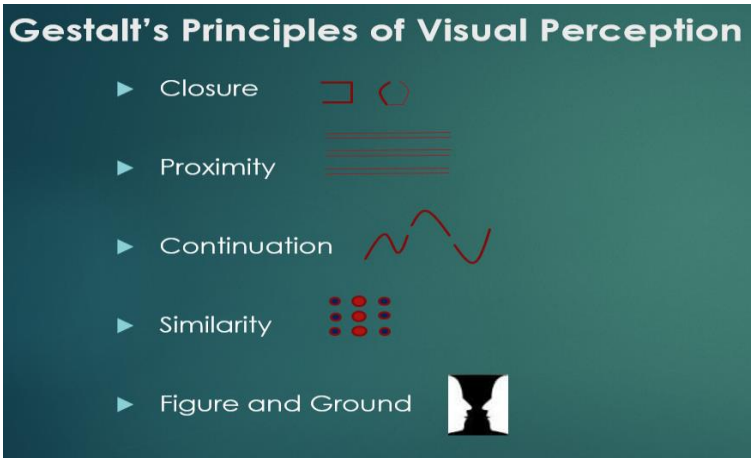

Figure 3.1.

[4]"The Gestalt Principles," accessed 10/04/2018, http://graphicdesign.spokanefalls.edu/tutorials/process/gestaltprinciples/gestaltprinc.htm; Kurt Koffka,"Gestalt Principles," Interaction Design Foundation, accessed 10/15/2018, https://www.interaction-design.org/literature/topics/gestalt-principles.

[5] Gale Thompson, "The 5 Principles of Gestalt," April 24, 2017, accessed 10/04/2018, https://sciencing.com/5-principles-gestalt-8430201.html.

Although visual acuity is not the same for all, Gestalt principles are a common human behavior. Think about a line chart applying the Figure and Ground principle. When dark and light shades are alternately applied would the audience view 1158 as a decrease or an increase? It is important to consider Gestalt principles to avoid conveying the wrong message.

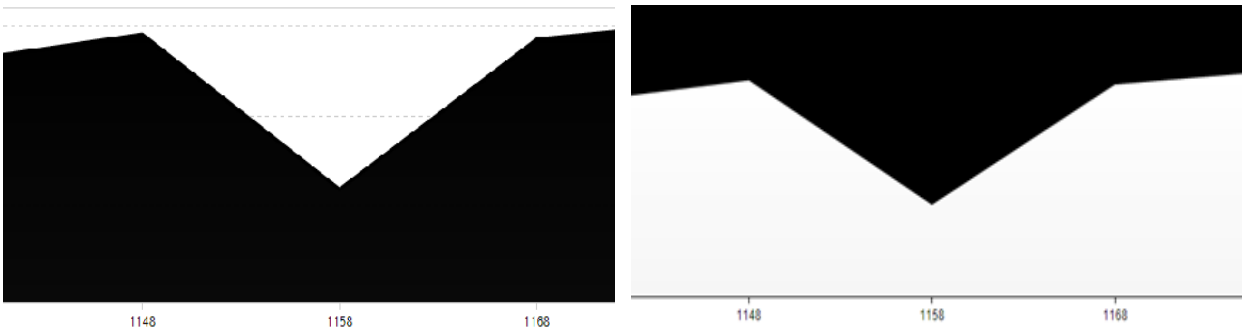

Figure 4.1.

Additionally, text objects can support visualizations. Items, such as titles, labels, legends and popup texts, are key components that add clarity and understanding.[6] Now let's apply the story into visualizations.

Presentation Types

Charts are a way to take information and make it understandable. The more information a chart is able to convey without increasing complexity, the better.[7] The tricky part is determining the visual image that simplifies interpretation and draws conclusions without misrepresenting the data. Therefore, a good chart will accurately convey information as a part of a larger whole, its context. Context is what provides its relevance.[8] Deciding the best way to represent data is a matter of selecting the best graphic in terms of the data's context. There are many design tools to assist with this process, such as Jonathan Schwabish and Severino Ribecca's *The Graphic*

[6] Bill Shander, "Data Visualization for Data Analysts," Lynda.com, accessed https://www.lynda.com/Excel-tutorials/Data-Visualization-Data-Analysts/178123-2.html?srchtrk=index%3a1%0alinktypeid%3a2%0aq%3aData+Visualization+for+Data+Analysts+with+Bill+Shander+%0apage%3a1%0as%3arelevance%0asa%3atrue%0aproducttypeid%3a2.

[7] Walter Hickey, "The Worst Chart in the World," *Business Insider* (June 17, 2013), accessed 09/21/2018 https://www.businessinsider.com/pie-charts-are-the-worst-2013-6.

[8] Antony Unwin, "Good Graphics," *Handbook of Data Visualization* (Heidelberg: Springer Verlag, 2008), 58.

Continuum as well as presentation design consultant Andrew Abela's *Chart Suggestions – A Thought Starter*. Dr. Abela breaks charts into four basic presentation types: comparison, composition, distribution, and relationship.[9] The following discussion reviews the various chart types according to Abela's categories and considers best practices and uses.

Comparison Charts

Comparison charts examine two or more items, either among items or over time, to establish similarities and dissimilarities. Two common comparison diagrams are the line and column charts. A line graph is commonly used when the data is continuous, usually over time. Although column charts can also show data changes over time, the number of components should not exceed five.[10] During the spring 2016 and 2017 fish spawning seasons, state officials commissioned a study of the stock of fish in City Lake. In figure 5.1, a line chart and a column chart compare these spawns. The advantage of the line chart over the column chart, below, is clarity. When using a column chart to compare components over time, a higher number of time intervals on the x-axis cause the graph to be more imprecise and cluttered. In contrast, the line graph manages a high number of time intervals without obscuring the data.

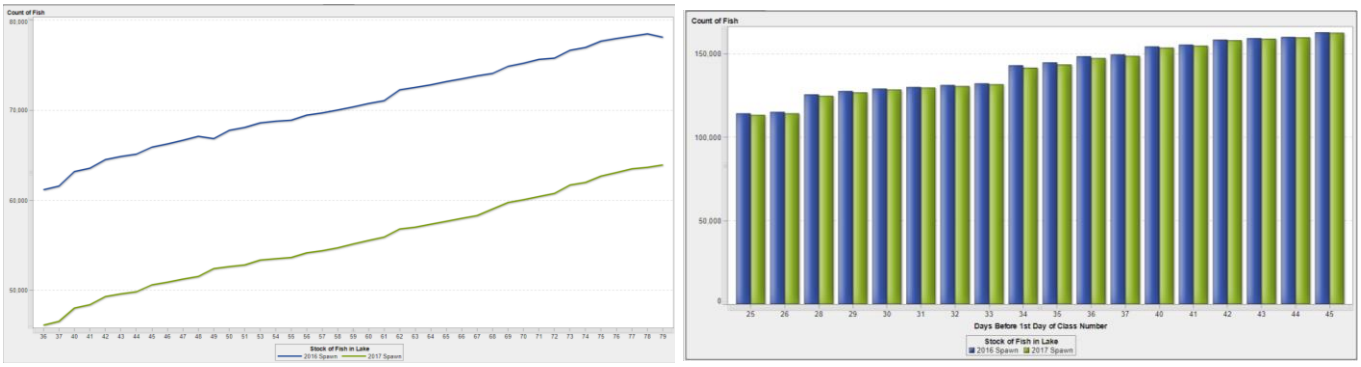

Figure 5.1. Comparison Charts - City Lake Spawns

When using a column chart, the numeric axis should begin at zero. According to Janis Gulbis, *Data Visualization – How to Pick the Right Chart Type EazyBI*, our eyes are sensitive to the

[9] Andrew Abela, "Choosing a good chart," *The Extreme Presentation Method* (September 6, 2006), accessed 09/21/2018 http://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html.
[10] Janis Gulbis, "Data Visualization – How to Pick the Right Chart Type," *EazyBI* (March 1, 2016), accessed 09/21/2018 https://eazybi.com/blog/data_visualization_and_chart_types/.

height of columns. Truncating columns can lead to spurious conclusions.[11] Figure 5.2 illustrates a comparison of the number of catfish to northern pike in City Lake. The un-truncated column chart on the left, accurately portrays the slight difference in the count of these two types of fish. However, the truncated column chart on the right visually may lead to the incorrect conclusion that the number of catfish in City Lake far exceed that of northern pike. Similarly, changes in the scale of a line chart can lead to wrong inferences. Although setting the axes at zero is preferable, it is not always the best representation of the data. The spring spawn of 2016 realized a thirty percent increase in the number of bass in City Lake. In Figure 5.2, the left line chart is auto-scaled by SAS Visual Analytics (VA), while the right line chart is manually scaled to begin at zero. In the graph auto-scaled by SAS VA, the change in the number of bass, although significant, appears better than thirty percent. On the other hand, when the y-axis is set to start at zero, the change in the amount of bass appears insignificant. To obviate these perceptual problems, the use of line markers and labels improves understanding.

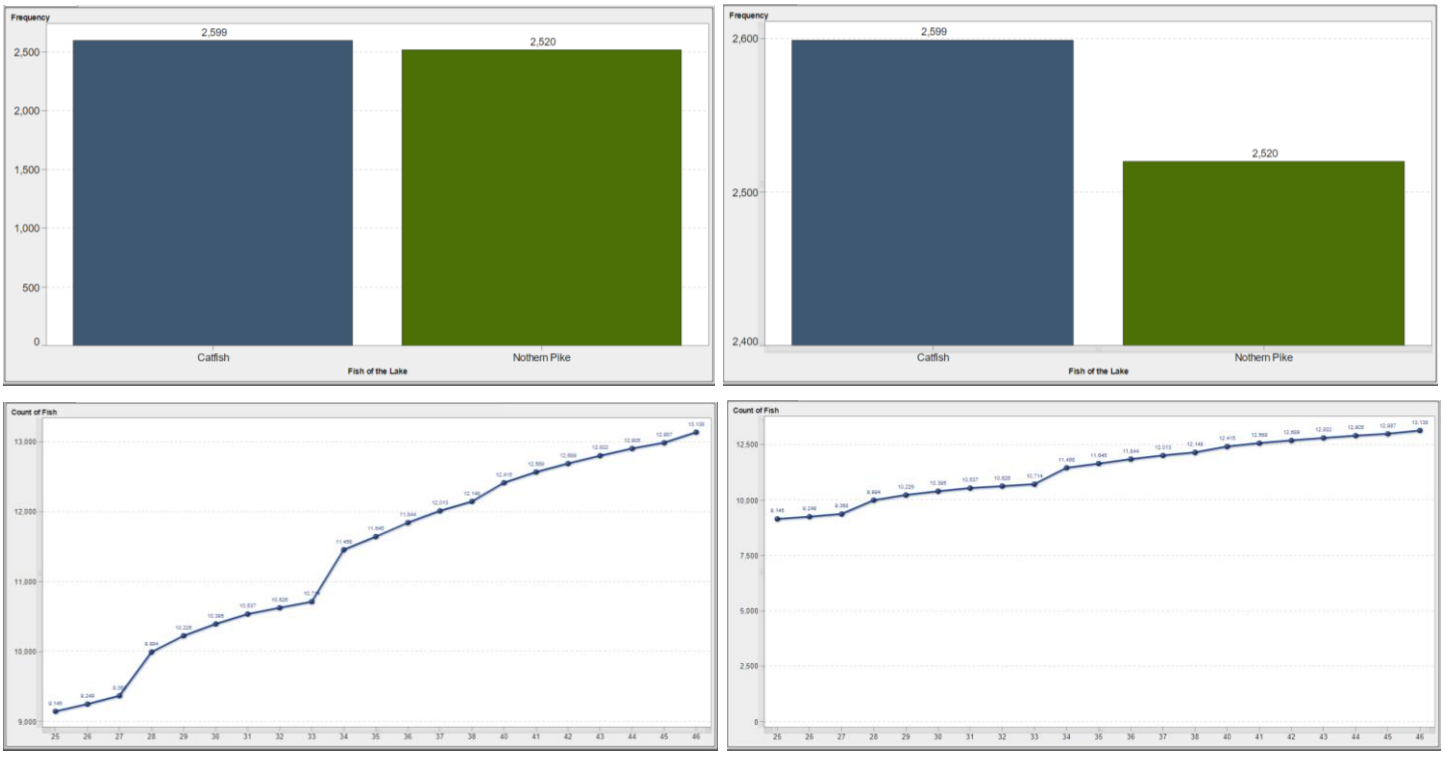

Figure 5.2. Comparison Chart - Column Chart Truncation

[11] Janis Gulbis, "Data Visualization – How to Pick the Right Chart Type," *EazyBI* (March 1, 2016), accessed 09/21/2018 https://eazybi.com/blog/data_visualization_and_chart_types/.

A good chart will convey information, but no amount of design can bring meaning to a graph in which its components lack relevance. A red-flag here, is that these comparisons often require a greater degree of design finesse to force the association of incompatible components. The dual axis column-line chart in figure 5.3 compares a school's fall student enrollment numbers to the percent of the city's squirrels found on campus. Even though the graph implies that higher enrollment causes fewer squirrels to be on campus, squirrels are not a subset of the school's total enrollment. As a result, the graph similarly suggests that a percentage of the school's fall students are squirrels. Accordingly, the comparison lacks relevance. In the absence of contextual integrity conclusions become problematic.

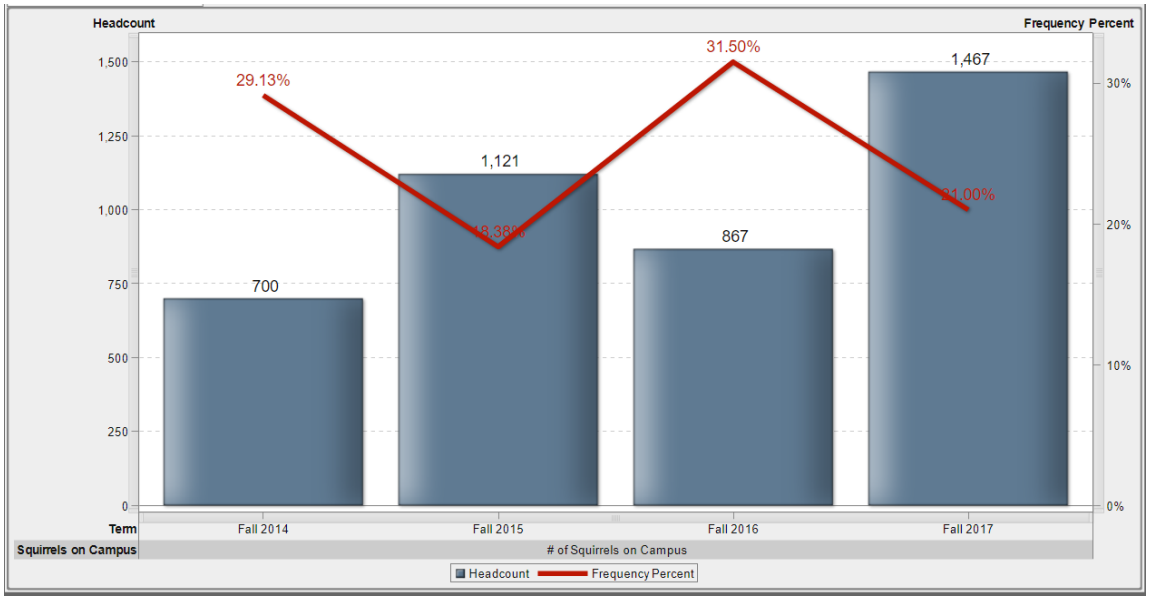

Figure 5.3. Are Squirrels Relevant?

Composition Charts

Composition graphics, such as pie and waterfall charts, visualize a part to its whole. Pie charts are frequently avoided because they can make it difficult to compare relative amounts. Actually, this type of graphic is best used to compare an amount to the whole as opposed to other components of the pie. One of the most common mistakes is the use of multiple components. In figure 6.1, the left pie chart represents the various types of fish in City Lake. The many kinds of fish that comprise the graphic diminishes the representation of the data, makes conclusions imprecise, and causes the chart to appear messy. The pie chart on the right prioritizes certain fish, thereby adding clarity to the components' relationship to the whole – not to each other.

Alternatively, a waterfall chart is a form of composition data visualization that shows the cumulative effect of subsets of records that may increase or decrease. As opposed to the pie chart, the waterfall chart allows for a better understanding of the relationship between individual components and as a part of the whole. The waterfall chart in figure 6.1 displays the same set of data used in the adjacent pie charts. Because waterfall charts employ straight lines to guide sight from component-to-component-to-whole it is better suited than the pie chart to show the relationships between individual components. Even then, the waterfall chart makes it difficult to determine if bass comprise fifty percent of the total fish in City Lake. A quick glance at the adjacent pie chart, it is easy to determine that bass make up less than half the total number of fish, a conclusion not apparent in the waterfall chart. Ultimately, the pie chart is limited to comparisons of singular components to the whole, while waterfall charts add clarity to evaluations of sub-sets to each other.

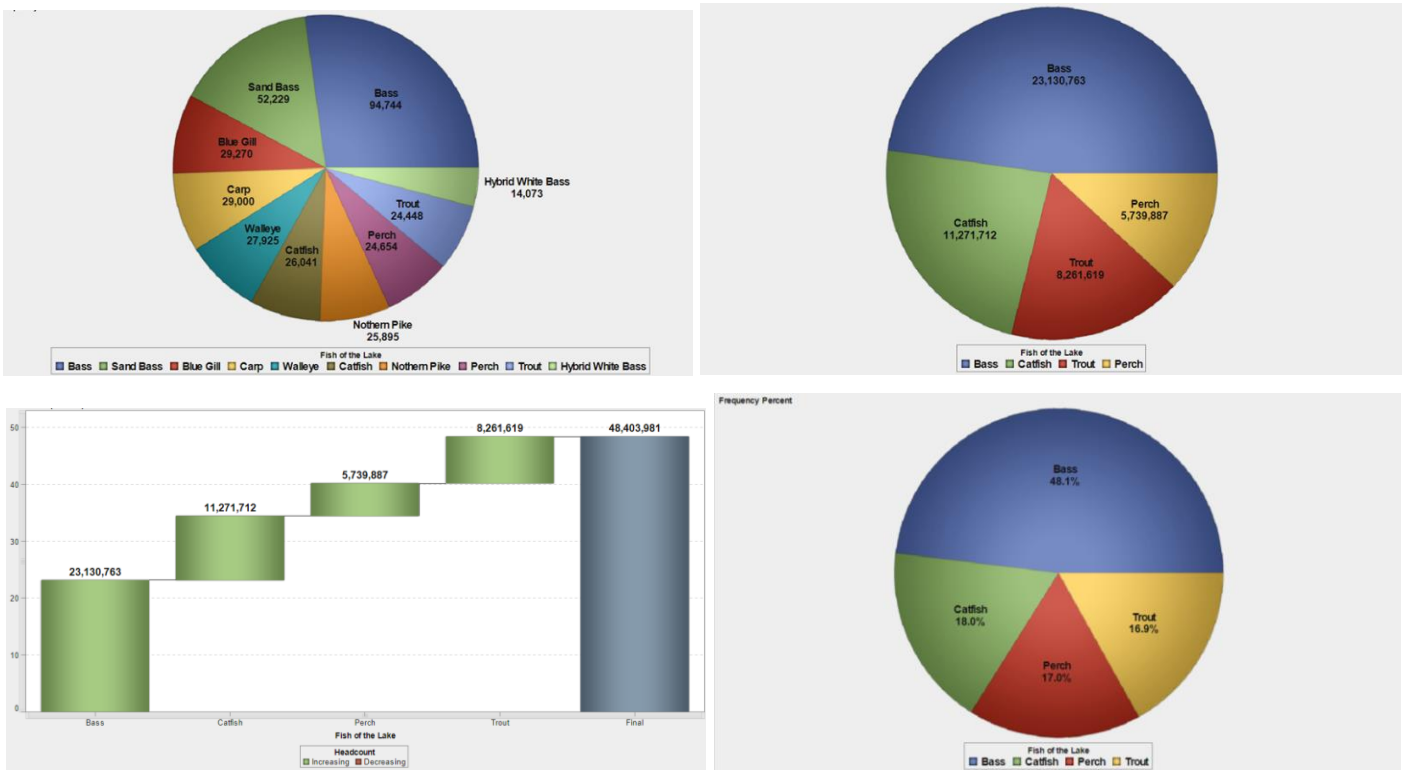

Figure 6.1. Composition Charts

Relationship and Distribution Charts

A Treemap displays hierarchical data by nesting figures, usually as rectangles. In figure 7.1, each component is assigned a box, which contains its nested internal subcategories. When a quantity is assigned to a category, its area size is displayed in proportion to that quantity and to the other quantities within the same parent category in a part-to-whole relationship.[12] Like a waterfall chart, a treemap shows the relationship between components both in terms of its portion of the whole and to other components, and adds a third component – frequency represented by shades of color.

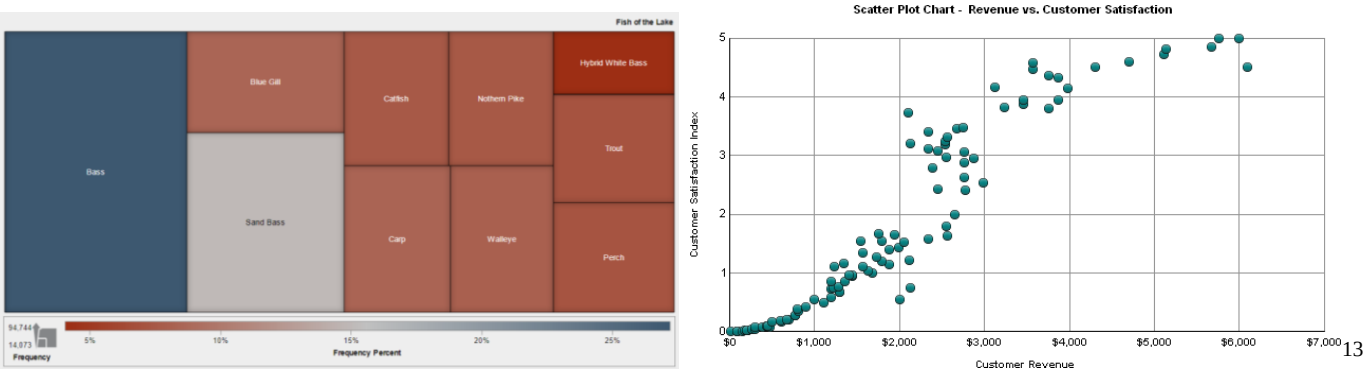
[13]

Figure 7.1. Treemap and Scatter Plot

In figure 7.1, a scatter plot is a two-dimensional data visualization that uses dots to represent the values obtained for two different variables - one plotted along the x-axis and the other plotted on the y-axis. Scatter plots show the relationship between two components and the variables' correlation.[14]

Color

Visualizations cannot communicate great stories without color, however there is a thing as too much color, not enough color, and simply the wrong colors. The first step to color is to identify the firm's brand colors through the company's website or through the marketing department.

---

[12] "Treemap," *The Data Visualization Catalogue*, accessed 10/09/2018, https://datavizcatalogue.com/methods/treemap.html.

[13] "Scatter Plot," microstrategy.com, accessed 10/12/2018, https://www2.microstrategy.com/producthelp/10.6/AdvancedReportingGuide/WebHelp/Lang_1033/Content/AdvancedReporting/Scatter_plot.htm.

[14] "What is a Scatter Plot and When to Use It," *Chartio*, accessed 10/09/2018, https://chartio.com/learn/dashboards-and-charts/what-is-a-scatter-plot/.

Identify the hex color numbers or whatever the analytical tool uses to associate certain colors. If the organization does not have a specific color palette, then find websites to assist building a color spectrum such as Adobe Color CC, Paletton, or Canva. Fuse colors together to see what works for the visualization. However be aware of printing with a black and white printer when utilizing color and if it is color blind friendly. Certain colors can be switched for people with color deficiency such as red and green or blue and purple. Also most organizations can associate negative impacts with the color red and green to a color blind person can mean the same thing. Obtain applications or websites to aid designing visuals which are color blind friendly like Color Blind Pal[15] which shows the colors a color blind person sees and also helps in the ability of seeing the true colors for people who have a color deficiency.

Conclusion

We have given the basic tools to assist in developing great visualizations. Review requests from the audience and identify key questions to answer with the data. Research, explore and deep dive the data identifying every possible angle. Utilize a storyboard to outline the story and remove any unnecessary data elements. Be aware of visual perception and don't deceive the audience. Labels, text, legends and colors bring visualizations to life and move along the storyline. Remember to use the correct chart or graph that will display the data effectively. Think outside the box with a treemap or scatter plot when designing for relational associations. Design appropriately in regard to what is being shown and always look for outliers within the visualization. Remove the mess and elevate to the best with these simple techniques.

Additional Reading

Chen, Chun-hou, Wolfgang Härdle and Antony Unwin. *Handbook of Data Visualization*. Berlin: Springer-Verlag Berlin Heidelberg, 2008. DOI 10.1007/978-3-540-33037-0.

Underwood, Jen. "Learning the art of analytical storytelling." *SAS* *https://www.sas.com/en_us/insights/articles/business-intelligence/learning-the-art-of-analytical-storytelling.html*.

[15] Color Blind Pal, colorblindpal.com. Color Blind Pal for Android and iOS helps people who are color blind see the colors around them.

About the Authors

De'Ambra Unuigbey

De'Ambra Unuigbey is a Business Intelligence Analyst at the University of North Texas (UNT) in Denton, TX. She has worked with SAS Visual Analytics and is certified in SAS Visual Statistics: Interactive Model Building. Prior to joining UNT, she worked as an Operations and Business Service Analyst at JP Morgan Chase within the Corporate and Investment Bank. De'Ambra has a BA in International Business and Spanish and an MBA in Business.

Steven Collins

Steven Collins has eleven years of experience in higher education, which includes three years as a Business Intelligence Analyst at the University of North Texas (UNT). He is a doctoral student at UNT, an Adjunct Instructor of U.S. history for North Central Texas College, and has a BA in Business and Management and an MA in European History.