A Survey on Types of Data Visualization Techniques using Behavioral Risk Factor Surveillance System Dataset

Prasad Bhoite
Panther id - 5488586

Poonam Gupta
Panther id - 6106375

Mohammad Asif Khan
Panther id - 6231979

Abstract

With the rapidly growing scale at which data is being collected today, reviewing raw data is humanly impossible in a reasonable amount of time. As our minds are more receptive of visual cues and interpret visualizations better than text or numbers, Data Visualization techniques are often utilized not only to review data for initial exploratory analysis but also for presentation of results and conclusions in a way that can be easily communicated to the target audience. In this paper, we survey a number of tools and libraries that can be used along with the R programming language to analyze the data, types of visualization techniques and corresponding examples using Behavioral Risk Factor Surveillance System (BRFSS) dataset. Finally, we review factors to consider when leveraging the various Data Visualization techniques.

Keywords Data visualization, trends, visual analysis, visual presentation, visualization tools, libraries.

1 Introduction

As the amount of data we collect and store is growing, the analysis of data, identifying patterns, trends are becoming increasingly difficult. According to [9] published in 2009, the researchers of the University of Berkeley estimated that the quantity of information in the world has increased approximately two exabytes every year. However just 7 years later, in 2016 IBM [2],updated this to 2.5 exabytes (or quintillion bytes) per day. This creates challenges for us to process, interpret and analyze this large volume of data. Sifting through this raw data is humanly impossible. A number of studies have been conducted [20],[18] and there is consensus that human minds respond to and process visual data better than any other type of data. In fact, the human brain processes images 60,000 times faster than text, and 90 percent of information transmitted to the brain is visual [20],[18].

As described in Oxford dictionary, 'visualize' can be defined as: "To form a mental vision, image, or picture of (something not visible or present to the sight, or of an abstraction); to make visible to the mind or imagination.". Similarly, Data Visualization or Data Viz has been described as the graphic representation of data and information, in a way that communicates clearly relationships to the observer in a meaningful or useful way. The goal of Data visualization is to leverage the cognitive ability of visual perception of the human brain to present data in a condensed way to understand information in such data. A good chart, graph or picture of data helps the user not only to remember information but also to quickly locate 'information', 'patterns' as well as anomalies from the data. The old english adage 'A picture is worth a thousand words' is certainly applicable and entirely fitting here.

As authors Liu et al. highlight in [7], today data visualization is widely used not only in the specialized fields of scientific research like health, manufacturing, sales, business systems, but also in the fields of education, sports, social media, ballot data, student histories, image and videos etc. In the current information age, data visualization is also used by average users for deciding on which products or services to buy, or as demonstrated in the current ongoing crisis of Covid-19, to review and keep up to date with health data around trends in infections, mortality rate etc.

Authors Ltifi et al.[9], propose that static visualization has a certain value for a user. However, due to the curious nature and the ability of humans to analyze data from different points of view, greater value can be derived if these visualizations can be interactive, giving the user the ability to zoom, pan, filter etc. They state "to solve this problem of interpretation, it is thus from now on necessary to develop interactive and centered user techniques, of visual exploitation of data in order to increase the improvement of confidence and comprehensibility of the model (generated by the KDD process) and the possibility of using the human capacities in pattern recognition."

Data visualization not only presents opportunities for data exploration but also conveying trends, patterns that can be used in making important decisions as well as communicating results.

To aid this, various data visualization techniques have been developed over time. In this paper we have reviewed a few of them using the dataset provided by Behavioral Risk Factor Surveillance System (BRFSS). BRFSS is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. We have used the most recent dataset from 2019.

Here is an example of data from this dataset.

This table shows the status of health coverage by education level. The healthcare coverage increases with the education level.

Education Level	**Yes**	**No**	**Missing**	**Refused**	**Not asked**	**Total**
Never attended school or only kindergarten	366	250	2	1	0	619
Grades 1 through 8 (Elementary)	6788	3071	57	24	0	9940
Grades 9 through 11 (Some high school)	15411	3928	119	48	0	19506
Grade 12 or GED (High school graduate)	98197	12822	536	332	0	111890
College 1 year to 3 years (Some college or technical school)	106863	9234	293	199	2	116591
College 4 years or more (College graduate)	151438	6111	162	176	0	157887
Refused	1443	249	23	94	0	1809
Not asked	18	4	0	0	4	26
Total	380524	35669	1192	874	9	418268

2 Survey

2.1 Introduction to tools and libraries in R:

A number of tools and libraries have been developed to aid data visualization. Here are some packages available in the R programming language.

R language has a built-in graphics package that allows creation of base graphs such as scatter plot, box plot etc however per [6] this is fairly limited as we need to design complex layouts or more control of layout is desired, this falls short.

Lattice Graphs: Lattice package is an improvement over the built-in R Graphics package that uses Trellis graphs and is used to visualize multivariate data. Trellis graphs enable examination of complicated relationships between multiple variables. [17]

ggplot2: It is a data visualization package for the statistical programming language R. It was created by Hadley Wickham in 2005. This is an implementation of Leland Wilkinson's Grammar of Graphics which is a general scheme for data visualization that breaks up graphs into semantic components like scales and layers. [11] provides a great comparison between base graphics package, Lattice and ggplot2. The base library produces static images and is not interactive. There are additional packages like gganimate [10] that add the layer of animation.

Plotly: Plotly on the other hand adds the layer of interactivity and animation on top of ggplot2 in R. This is based on a JavaScript library 'plotly.js' and uses JavaScript Object Notation (JSON) specification to represent, serialize and render web graphics. This enables creation of interactive, high-quality, publication-quality graphs. We can use this library to make a variety of graphs including subplots, heatmaps, multiple-axes and 3D (WebGL based) charts. We can use the htmlwidgets package in RStudio to leverage the framework it provides to bind R commands to various, interactive JavaScript libraries including ones that generate data graphs such that the data analysis can be done at R console or be saved as standalone web pages. The htmlwidgets framework also enables rendering of graphics locally.

Spatial/Geographic data Depending on the type of data, there are specialized packages available. One such package for visualizing geographic data (which is a type of spatial data) is ArcGIS. This software package allows users to quickly create maps by using geographic knowledge [23].It is a proprietary software offered by the Environmental Systems Research Institute (ESRI), which is considered the world leader in Geographic Information Systems (GIS).When working with geographic data, there are two factors that establish the tools and techniques that can be used for data visualization of such data - the format of data and the presentation of this data over maps [3]. A number of open and proprietary formats exist. For our survey, we used ArcGIS Pro. We also show how to use choropleth maps using Plotly.

2.2 Types of Visualization techniques:

2.2.1 Temporal Visualization:

A Data visualization falls under a temporal category if they satisfy the following two conditions:

- It represents linear plots.
- It is one-dimensional.

This type of category consists of lines that are stand-alone or overlapped having a starting and ending time and also track time series data over a period of time. Some of the visualization techniques in this category are:

a. Scatter plot: It's a type of plot that uses dots to represent values of two different features on a 2-d plane. The position of the dot with respect to the vertical and horizontal axis represents the value of the data point. It helps us determine the correlation between the two features. Scatter plots have multiple design variants which can be used to display the data in case of different datasets [16].

b. Line Chart: This type of visualization technique displays data in the form series of data points marked and connected through segments of straight lines, forming a continuous line across the plane. Line charts are often used to visualize the data over a period of time/ time series data. The line chart is equivalent to scatter plot in a sense. However, measurement points are ordered often with respect to the X-axis and the points are connected via segments of straight line.

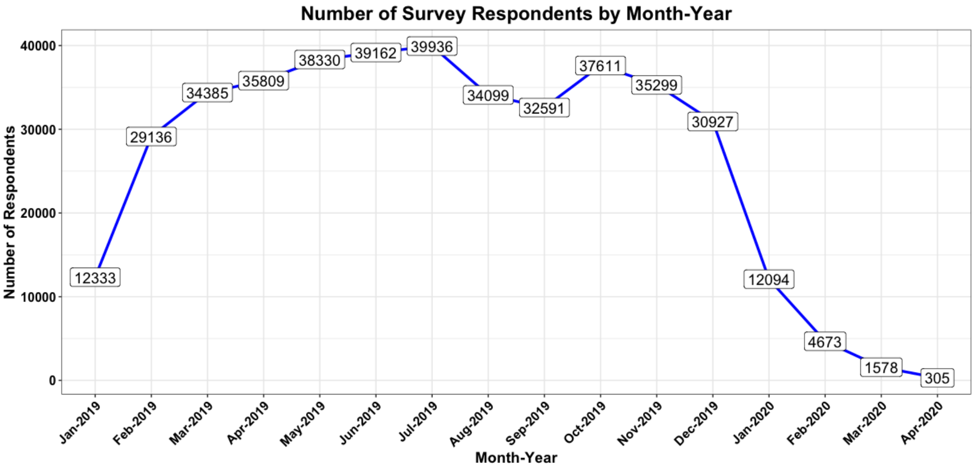


Figure 1: Number of survey respondents by month-year from January 2019 through April 2020

c. Polar area diagram: This type of visualization technique helps represent time series data in a cyclical way. It resembles a pie chart, however the sectors in the diagram differ by angles as well as how far they extend from the center of the circle. It can be used to display multiple features' information of a dataset stacked on one another in a single pie. The data generated due to the cyclic phenomenon are ideally represented by this visualization.

d. Bar Chart: Bar charts are used to represent data in terms of vertical or horizontal bars. The value of the bar determines the value of a feature at a point of time. This type of chart helps compare patterns of change of a feature/variable over a particular period of time [15]. The bar charts can be used to compare groups of data for a particular variable over time in terms of grouped or stacked charts.

e. Stacked Line Graph: It's a more advanced version of line graph wherein multiple lines are stacked one over another and it can be used to compare and get a much clearer picture of associated data/lines/categories in the dataset

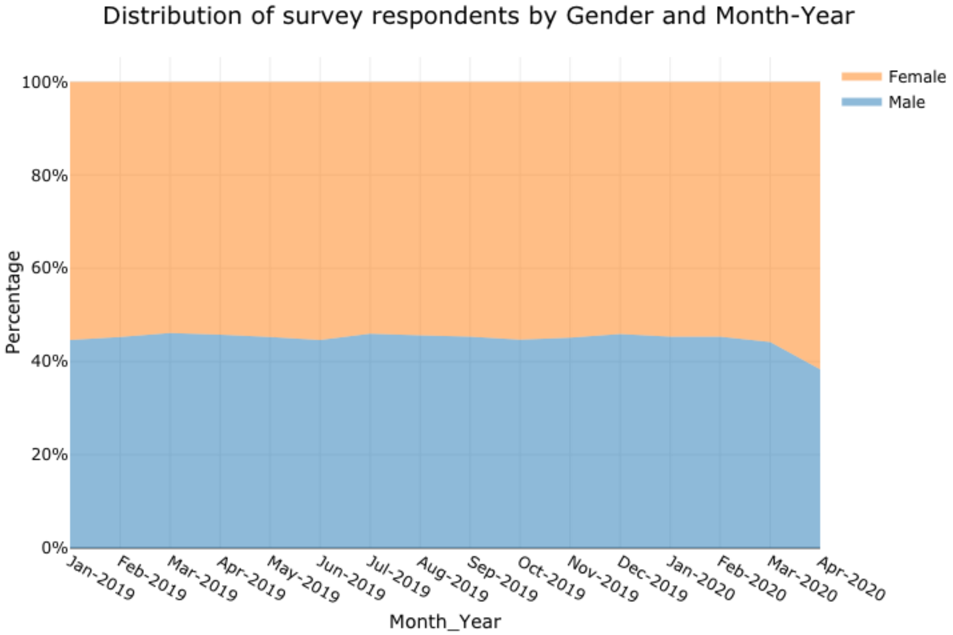


Figure 2: Stacked Line Graph (Male and Female survey respondents Jan 2019 through Apr 2020)

d. Gantt Chart: Gantt chart is the visual representation of work completed in terms of horizontal bar chart in a certain period of time with respect to the allocated time. This visualization type is heavily useful in project management tasks. It is helpful for distributing the project into various tasks and they can be spread across different timelines. We use different colors and varying placement of bars to help us determine resources used across different tasks and the time it takes to complete.

2.2.2 Hierarchical Visualization:

This category describes the relationship between the items with other items in terms of parent-child hierarchy relationship. This type of visualization displays the ranks between items, distributed data among different items in a tree like structure. Some of the following techniques fall under this category.

a. Tree diagram: This technique is the most basic of all coming under the hierarchical category. As it's intuitive from the name itself, it displays the items/object in the form of a tree structure having a root which has no parents. All others except the leaf nodes have children. The tree diagram shows how items/nodes are correlated with one another and each node has some values. An example of a tree diagram is a management system displaying subordinates/superiors.

b. Treemap: Treemap, unlike the tree diagram displays the data in the form of rectangles. The entire rectangular area gets split into rectangles recursively in accordance to the hierarchy structure and the data attributes. The treemap can display the overall hierarchical structure with different sizes of rectangles depending on the attribute value taken in each rectangle[19].

c. Sunburst diagram: The sunburst visualization technique produces donut-like shape diagrams. The shape has different colors along it and the hierarchical data format is distributed just like treemap diagrams. However, the entire space is circular instead of rectangle. The space is further divided into subsections and each subsection has different size and color on the circular space based on its value.

d. Donut Chart/Ring Chart: As the name suggests, the chart is shaped like a donut or ring and is used to display the proportion of different components/features of data in the shape of a whole ring. The shape can be very much compared to a Pie chart if it's a single ring. However, It has an advantage over pie charts because we can use multi-ring donut chart for different data series.

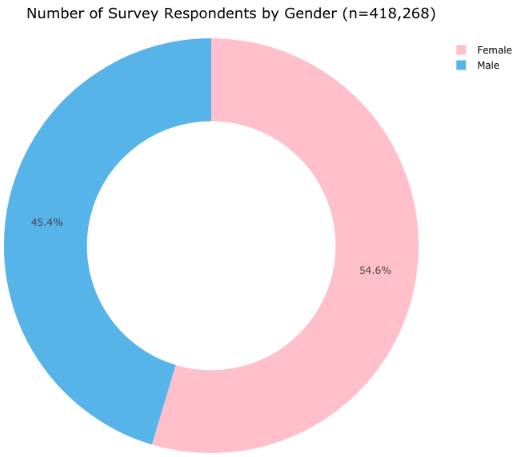


Figure 3: Donut Chart (Percentage of Survey Respondents, Female: 54.6 and Male: 45.4)

2.2.3 Network Visualization:

Just as the name suggests, network visualization displays the complex relationships among different elements in the data set. This kind of visualization displays the directed and undirected graph diagrams. The following visualization techniques come under network visualization.

a. Matrix chart: The visualization forms a matrix grid structure used to display the relationships among two or more variables/attributes of a data set. It's made of rows and columns and is an equivalent representation of cross tabulation. Each cell has color and a size. The color and density of the cell represents how correlated the row or variable is with another variable

b. Node-link diagram: Node- link diagrams are just like the graph diagrams which can be directed or undirected. It determines the relationship among different entities and the kind of relationship(Strong,Weak etc.) each node has with the other. It's widely used in our day to day life such as geo-referenced node-link diagrams or network diagrams wherein the nodes are placed in accordance with geographic criteria[4]. The visualization can further be divided into 3 types such as 1) Explicit vs Implicit 2) Directed vs Directed 3) Free, Styled or Fixed.

c. Word cloud: This visualization technique displays the words with different text sizes. The most often used words are displayed in larger texts as compared to the ones that are less used.This kind of visualization can provide good insights on comparing the usage of different words and find out the significance of a speech/letter/voicemail etc.

d. Alluvial diagram/Sankey Plot: Alluvial diagrams are a type of flow diagram originally developed to represent changes in network structure over time. In allusion to both their visual appearance and their emphasis on flow, alluvial diagrams are named after alluvial fans that are naturally formed by the soil deposited from streaming water.This enables us to get a distribution flow of some numeric value by multiple categorical variables.

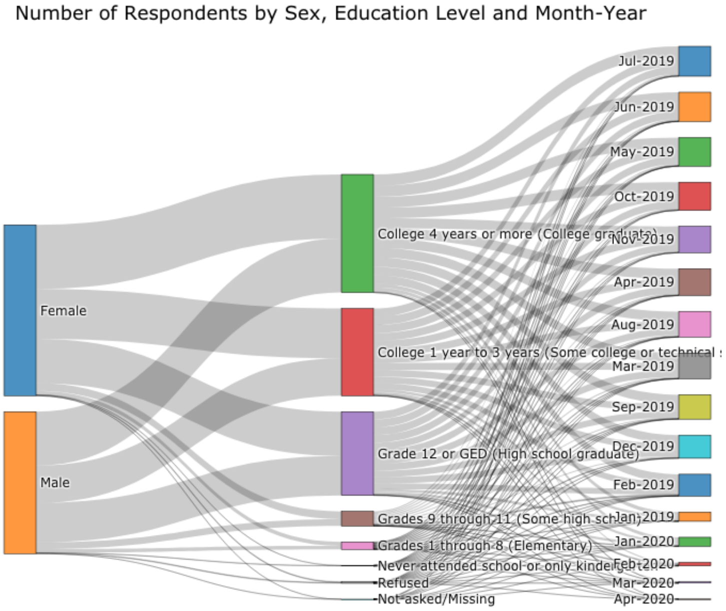


Figure 4: Distribution of survey respondents during 2019 by gender, education status, and month-year

2.2.4 Multidimensional Visualization:

Just as the name suggests, multidimensional visualization has multiple dimensions. It uses more than two variables or features of a dataset to create a graph of visualization. These visualizations are visually appealing and eye-catching. The following techniques fall under multidimensional visualization.

a. 3D Glyph Plot: Glyphs are mainly used to display multidimensional data. This plot helps us visualize large complex datasets in the form of glyphs. Glyphs are small shapes made to visually identify different categories in the data according to its shapes[8].

b. Stacked Bar Graph: This kind of graph shows the data in the form of bars and delineates a comparison between different categories of data. The complete bar displays whole data with regards to a particular entity and sections/segments in the bar are the different categories of the whole bar. This data can be both represented in the form of 2-d and 3-d visualizations.

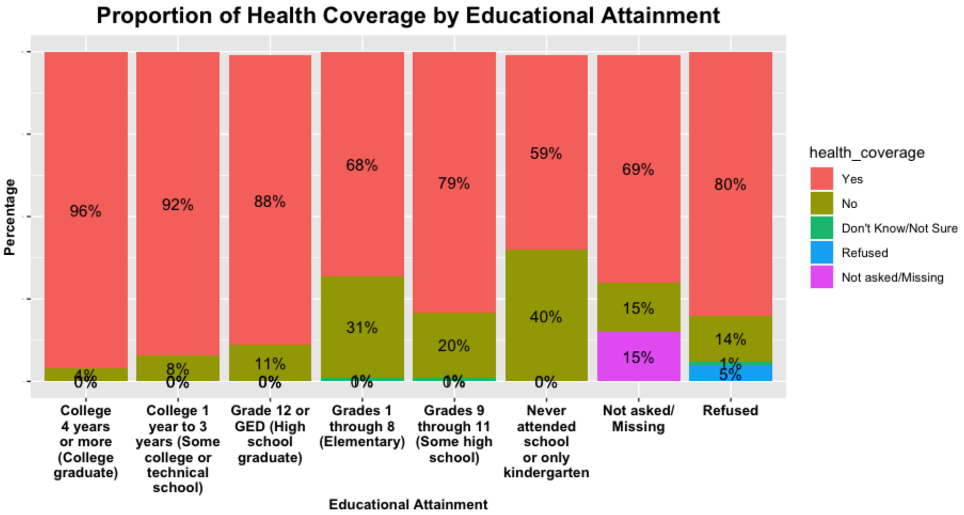


Figure 5: Stacked Bar Graph (Proportion of Health Coverage by Educational Attachment)

c. Histogram: Histograms are representation of data or values in terms bars. Unlike bar graphs, histograms represent data of a particular range of values. Histograms are good for understanding the distribution of data values and it also represents the outliers in the range if there's any. The histograms represent a continuous range of values and their corresponding data value in the graph. The flatter histograms tend to show less variability than the bumpier histograms[5]

d. Pie Chart: Pie charts are widely used visualization techniques that show the proportion of different categories by dividing the entire circle into different sections. It's easy and gives a quick view of the proportion of different categories. However, it has multiple downsides such as It's not ideal for large size of data. It takes up a lot more space as compared to many other visualization techniques.

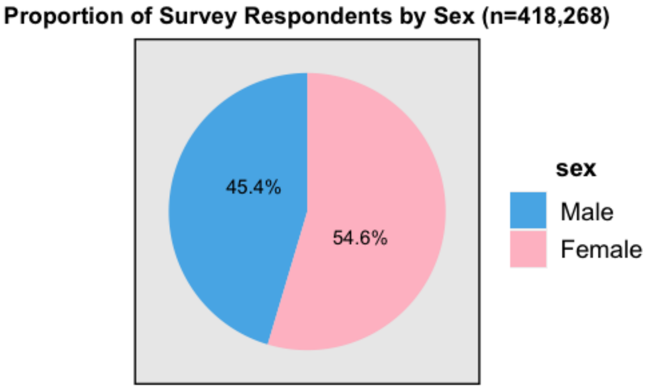


Figure 6: Stacked Bar Graph (Proportion of Health Coverage by Educational Attachment)

2.2.5 Geospatial Visualization:

Geospatial data Visualization is widely used to provide visual representation of the geographic location data with respect to any statistics [12]. This type of visualization technique has innumerable applications namely marketing, finance, e-commerce, advertising, education, telecommunications etc. and is very helpful in the field of data science to make data driven decisions. The visualization techniques such as Flow Map, Density Map, Heat Map, Choropleth etc. come under Geospatial Visualization.

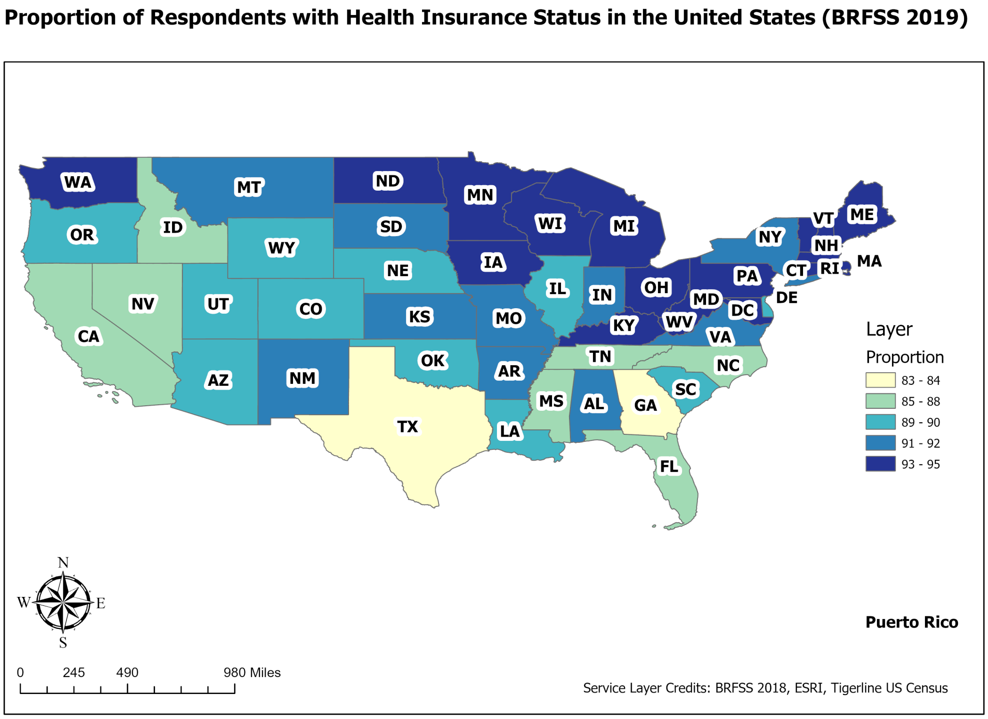
a.Flow Map: This Map can be considered as a hybrid of flow diagram and map. It uses linear straight/curved/dotted lines to represent the movement over a particular geographical area in the map. This is extensively used in multiple applications such as displaying species' migration, money flow, travelling, trades, ideas, road traffics, telecommunications etc.

b.Density Map: As the name density suggests this map helps us identify the density of a particular feature data over a geographic region in the map.They are highly useful to determine concentration of a feature data over a given geographic region, few of the applications include disease/death/birth density of a region.

c.Heat Map: Heat map is a highly useful data visualization technique used to the magnitude of feature data/phenomenon in terms of colors over a 2-d space. It can be used over large and complex datasets to make an easy interpretation. Some applications of heat maps are big data analytics, website tracking etc.

d.Choropleth Map: This type of visualization maps display the statistical data through various color patterns or symbols on the geographic region (Continent/Country/State etc.). These maps are ideal for recognizing the variability of different regions regarding a particular statistics in the map with respect to the color patterns/symbols. One of the downsides of this map is , the regions with bigger areas are likely to have bigger and better interpretation as compared to a smaller region.

Here is the choropleth map for data from BRFSS dataset using ArcGIS Pro


Proportion of Respondents with Health Insurance Status in the United States (BRFSS 2019)

Service Layer Credits: BRFSS 2018, ESRI, Tigerline US Census

Figure 7: choropleth map for data from BRFSS dataset using ArcGIS Pro

Here is the choropleth map for data from BRFSS dataset using plotly library.

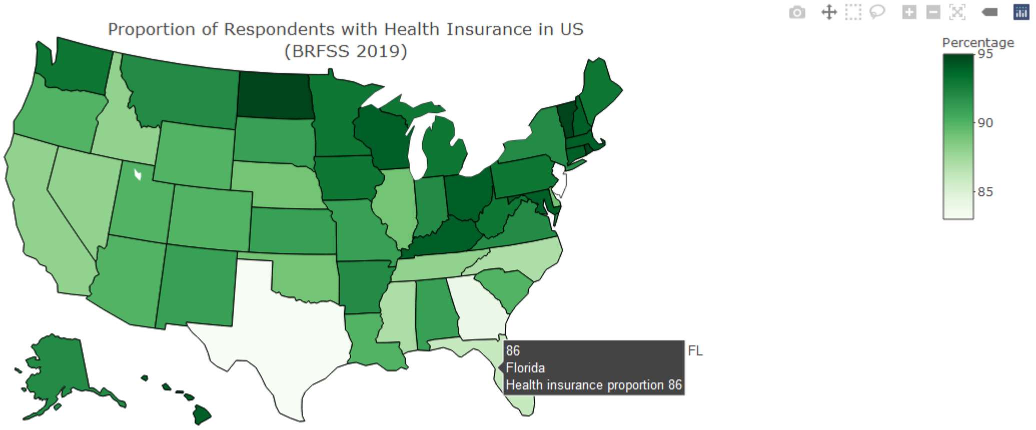

Proportion of Respondents with Health Insurance in US
(BRFSS 2019)
Percentage
95
90
85
86
Florida
Health insurance proportion 86
FL

Figure 8: Choropleth Map (Proportion of Respondents with Health insurance in the US in 2019)

3 Factors to consider

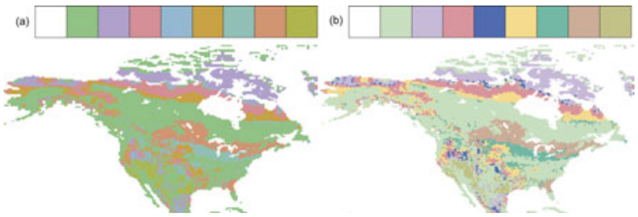
As we see above, a single graph or picture can show important information that can span 1000's of pages of raw data. However, not all visualizations are created equal or even appropriate for all types of data. With such a large number of tools and techniques, we list below some important considerations we found during our survey.

Human perception and amount of data: As the complexity of data increases, especially when the number of factors, human eyes can have difficulty in extracting meaningful information when data becomes extremely large. [1]

Screen size: As the amount of data increases, identifying the appropriate visualization technique and scale becomes critical, as we have limited screen size. If not designed carefully, this could result in visualizations that are too dense for any useful information to be derived from and may overwhelm the users.

Real time visualization: For data that is changing in real time, the speed at which a visualization is relaying the information is also critical in making real-time decisions; for example as in the case of the stock market data for trading where the delay in getting information and its visualization could make a big difference. Another factor to consider is if the tool, platform or software is able to keep up with the amount of data that is generated in a certain period of time.

Choice of color scheme: The color scheme used in a data visualization can have a significant impact on the information it is trying to convey. In [21], Ware summarizes that rainbow and grayscale colormaps are effective for metric and shape comprehension respectively. Not just that, depending on the visualization, the choice of colors in the colormap can affect the increase or decrease the clarity of information. In the example categorical colormap below, as shown in (a), is less than optimal whereas an improved version is visually optimized to show better contrasts and ultimately more useful to the observer.



Data visualizations should have a color plan that can relay the meaning of the elements they are trying to represent appropriately. For example, if we are representing a health condition that has severe consequences, showing that in green is not appropriate. Similarly, a visual representation of network traffic data may show a network outage in red. In addition, color should also be used consistently throughout the system since change in color pattern may signal a change in meaning.

Size/volume of data: As we mentioned earlier, the amount of data that is being generated today is very large. When data is pre-processed, cleansed and presented, the volume of data is a big consideration to see if the library, tool or software capable of handling such large volumes.

High dimensionality and complexity: As our ability to gather more data improves due to improvements in sensors, better definition and understanding of what data to collect, we are storing a large number of data attributes. With this growing number of attributes, there is a need for analysing the relationships between these attributes as well. This becomes another area where data visualisation can be leveraged for visualizing trends in multiple dimensions for both analysis and presentation of such data. The presentation of these multidimensional data visually can be confusing or overwhelming depending on the way it is presented.

Animation: Animation can bring data to life, during both the visual exploration and storytelling phases. It not only engages the viewers, it can also present 'information' in a meaningful way. This is especially true for multi-dimensional data. As an example, the users can see how data changes either over time or another variable and can watch the story unfold. The TED talk and BBC documentary [14] by Hans Rosling, one of the data visualization pioneers are two such prime examples.

However, as detailed by Robertson et al. [13] visualization using animation can also overwhelm the viewer or will need to review the animation multiple times before deriving the intended result or in some cases like if left on their own, the consumer may be lost in the transitions happening on the screen entirely. Instead they suggest using alternative methods that may be more appropriate in certain situations like using two static images that show incremental changes so users can actually see the difference between them and realize the full potential of the changes.

Interactive data visualization: Interactive data visualization can put the user in control over how to view the data and which aspects may need further 'visual' analysis by zooming in and out, panning or focusing in a certain section of data can be helpful. Interactive data visualization can enable users to sample, filter or even aggregate data. However this too, if not done correctly, can lose its value. As the users can rotate, zoom the display to interact with data, users may lose perspective by overzooming or overpaning the display. This not only happens with 2D visualizations however can get even worse with 3D visualizations. With 3D visualizations, when comparing various factors, as the data can be in different layers, the user may not see the relationships clearly.

Another issue is the latency and delays in the interactions. When the user interacts with the visualization and it takes a significant amount of time for the system to respond, which could be due to the amount of data processing or network latency. According to authors [18], even though the additional dimension may convey additional semantic information, due to complexity leading to additional cognitive demands and the lack of additional functionality and control, 3D visualizations lead to minimal advantages. According to authors Yu, Harrison and Lu [22], feature-driven storytelling animations with common interactive visualization techniques and conclude that feature-driven storytelling animations consistently perform better, timely results.

4 Conclusion

This paper provides a survey of a wide range of tools, techniques and libraries available to address various aspects of data visualization. Designing a system however that is both informative and aesthetically pleasing remains a challenge in this field. With recent advances in ability to utilize animations and interactive Data visualizations, these challenges have continued to grow. There are a number of factors that should be considered as part of a data visualization project such that the data exploration, data analysis and communication of results during the presentation phases can happen in a meaningful and effective manner.

5 Acknowledgments

The authors would like to thank Professor Antonio Hernandez for his direction and guidance for this survey paper.

References

[1] Rajeev Agarwal and Anirudh Kadadi. Challenges and opportunities with big data visualization.

[2] Michael Belfiore. How 10 industries are using big data to win big https://www.ibm.com/blogs/watson/2016/07/10-industries-using-big-data-win-big. *IBM*, 2016.

[3] Roger Bivand. Analysis of Spatial Data https://cran.r-project.org/web/views/Spatial.html. *CRAN*, 2021.

[4] A. Debiasi, B. Simoes, and R. De Amicis. Schematization of node-link diagrams and drawing techniques for geo-referenced networks. *International Conference on Cyberworlds (CW), Visby, Sweden*, 10:34–41, 2015.

[5] Kaplan, Jennifer J., John G. Gabrosek, Phyllis Curtiss, , and Chris Malone. Investigating student understanding of histograms. *Journal of Statistics Education*, 22(2), 2014.

[6] Nicholas Lewin-Koh. Graphic Displays and Dynamic Graphics and Graphic Devices and Visualization https://cran.r-project.org/web/views/Graphics.html. *CRAN*, 2015.

[7] Liu, Shixia, Weiwei Cui, Yingcai Wu, and Mengchen Liu. A survey on information visualization: recent advances and challenges. *The Visual Computer 30*, (12):1373–1393, 2014.

[8] Lombeyda and Santiago V. Distinct 3d glyphs with data layering for highly dense multivariate data plots. 2016.

[9] Ltifi, Hela, Mounir Ben Ayed, Adel M. Alimi, and Sophie Lepreux. Survey of information visualization techniques for exploitation in kdd. *In 2009 IEEE/ACS International Conference on Computer Systems and Applications*, pages 218–225, 2009.

[10] Thomas Lin Pedersen and David Robinson. gganimate https://gganimate.com/index.html.

[11] Roger D Peng. Plotting with ggplot2: Part 1 https://github.com/rdpeng/CourseraLectures/blob/master/ggplot2_part1.pptx. *Johns Hopkins Bloomberg School of Public Health*, 2013.

[12] Rhyne, Theresa Marie, and Alan MacEachren. Visualizing geospatial data. *In ACM SIGGRAPH 2004 Course Notes*, pages 33–es, 2004.

[13] G. Robertson, R. Fernandez, D. Fisher, B. Lee, and J. Stasko. Visual. comput. graphics. *IEEE Transactions on Visualization and Computer Graphics*, 14, 2008.

[14] Hans Rosling. Hans Rosling's 200 Countries, 200 Years, 4 Minutes - The Joy of Stats - BBC Four https://www.bbc.co.uk/programmes/b00wgq0l. *BBC*, 2010.

[15] Salkind and Neil J. Encyclopedia of research design. *Thousand Oaks, CA: SAGE Publications, Inc.*, 0, 2010.

[16] A. Sarikaya and M. Gleicher. Scatterplots: Tasks, data, and designs. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):402–412, 2018.

[17] Deepayan Sarkar. Trellis Graphics for R https://cran.r-project.org/web/packages/lattice/index.html. *CRAN*, 2020.

[18] Anne Trafton. In the blink of an eye. Retrieved from https://news.mit.edu/2014/in-the-blink-of-an-eye-0116. *MIT*, 2014.

[19] Y. Tu and H. Shen. Visualizing changes of hierarchical data using treemaps. *IEEE Transactions on Visualization and Computer Graphics,*, 13(6):1286–1293, Nov.-Dec. 2007.

[20] Vogel, Douglas Rudy, Gary W. Dickson, and John A. Lehman. Persuasion and the role of visual presentation support: The um/3m study. 1986.

[21] C. Ware. Color sequences for univariate maps: Theory, experiments and principles. *IEEE Comput. Graph*, 8(5):41–49, 1988.

[22] Li Yu, Harrison L, and Aidong Lu. Effectiveness of feature-driven storytelling in 3d time-varying data visualization. *Journal of Imaging Science and Technology*, 60(6), 2016.

[23] Liang Zhou and Charles D. Hansen. A survey of colormaps in visualization.