Relationship Types in Visual Analytics

Nur Izzati Shabdin
Advance Informatics Department
Razak Faculty Technology and Informatics
Universiti Teknologi Malaysia
n.izzatishabdin@gmail.com

Suraya Ya'acob
Advance Informatics Department
Razak Faculty Technology and Informatics
Universiti Teknologi Malaysia
suraya.yaacob@utm.my

Nilam Nur Amir Sjarif
Advance Informatics Department
Razak Faculty Technology and Informatics
Universiti Teknologi Malaysia
nilamnur@utm.my

ABSTRACT
Visual analytics are practically an application of the task of transforming the data into meaningful and reliable information in order to synthesize the knowledge about the data. Recently, organizations collect large data with aim of extracting useful information for organizational usage for better decision making. When the organization have a large and complex information, it leads to more complex relationship between the variables in the data. As a consequence, the visual analytics representation must be able to show, visualize and handle for more complex relationship of data. However, the research found the lack of relationship research in current visual analytics that lead difficulties to guide and design new relationship representation. Thus, this research aim to recognize, discover and categorize relationships types of visual analytics in representing a set of analytical data. Design Science Research Methodology has been used as research method for this study. It consists of two activities which are i). identify the visual analytic relationship context and its challenge and ii). analysis of relationship in visual analytics representation. At the end, this study is expected to identify and categorize the visual analytics representation according to six relationship types. This identification can help the visual analytics community to understand the primary and basic concept of relationship representation as a guidelines and knowledge for more comprehensive research in the future.

CCS Concepts
Human-centered Computing ~ Visualization ~ Visualization Application Domain ~ Visual analytics

Keywords
visual analytics; relationship; correlation; differences; comparison; distribution; outliers and reduction.

1. INTRODUCTION
In the organization, an analytical activity has become essential to translate the knowledge from raw data into a meaningful business decision. From managerial perspective, they need relatable data from various field to gain more comprehensive business decision. Therefore, relationship between analytical data to show the reasoning behind the interconnection is becoming more important to derive insights from various dataset. In recent years, the relationship become more significant when the organizations are starting to collect big data in more volume, variety and velocity. These large and complex data involves multi relationship between the variables and there is a demand for visual analytics to show complex relationship (e.g. more than two relationships).

This situation requires a better understanding of relationship in visualizing more complex of analytical data. The visual analytics community must be able to know and manipulate the theoretical of visual representation in order to present the relationship in more accurate and effective way. Perhaps, there is a need to develop new visual representation that able to represent comprehensive relationship or optimize the manipulation of current visual representation that available in the market. However, there is a lack of study about relationship representation for current visual analytics –thus, it creates large gap of knowledge in manipulating the relationship within current visual analytics or developing a new relationship representation. The current problems faced by analytical developer is the confusion between types of relationships and different visual analytics representation to be used in representing the data. Without categorization of relationship types, criteria and elements, it is difficult to choose an effective visual representation to represent the data since they do not understand how the elements are presented in order to show, clarify and strengthen the relationship between data. Therefore, the research found the importance to further digging more about relationship as it is essential for big data era. The identification of type, criteria and elements of relationship will become a basic research guidance for more insightful relationship in visual analytics.

ICCTA 2020, April 14–16, 2020, Antalya, Turkey

ACM ISBN 978-1-4503-7749-2/20/04…$15.00
https://doi.org/10.1145/3397125.3397127

2. BACKGROUND OF RELATIONSHIP IN VISUAL ANALYTICS
Analytics are practically an application of the task of transforming the data into meaningful and reliable information in order to synthesize the knowledge about the data [1]. To facilitate an analytical task, visual analytics is needed in addressing the problems such as size, complexity and some cooperation between human and machine analysis to prevent it from unmanageable. By focusing on analysis condition - visual analytics has been define as the combination of automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets [2]. In complex information, visual analytics has the capacity in helping the analyst to explore and evaluate the large amount of data [3]. Visual can help to show several varieties of unseen trends, patterns and unknown correlation that may exist in the data [4].

Furthermore, [5] emphasized the needs of interactive visual interface to support analytical reasoning process. Harman [6] described reasoning as an intellectual, psychological or internal process of changing individual's view and perceptions while Walton [7] portrayed reasoning as a linguistic interaction process that seems to be more sociological than psychological. In relation to these two contexts, analytical reasoning is a fundamental for visual analytics as it applying human judgement in achieving a conclusion from an integration of solid evidence, presumptions throughout analysis process. The research found that visualizing the relationship is one of the main elements to facilitate the reasoning process. This happen when the analyst has to compare two or more sets of datasets. An effective visual representation must be able to help the analyst to discover the relationships, differences and similarities of the variables among the datasets. Clarity on the relationship will induce the reasoning, judgement and justification during the analysis process.

When representing complex data, the number of variables represented in the data becomes significantly large, then the relationship between the variables will be difficult to interpret and manage.According to [9], the relationship in visualization can be defined as representation of visual observable interactions between subject and object. It is the way in which two or more things are connected, or the state of being connected and being describe and visualized in a particular way. Relationship is important to show the interconnection of data between various field. From the previous study, [25] has identified six categories of relationships in visualizing the data: i) correlation, ii) distribution, iii) comparison, iv) difference, v) outliers and vi) dimension reduction. Each of the relationship type has been defined as Table 1.

Table 1. Six Types of Relationships in Visual Analytics

Type of relationship	Operational Definitions
Correlation	A relationship that exist between variables and can show how strongly related they are.
Comparison	An examination of two or more items to establish similarities and dissimilarities.
Distribution	The position, arrangement, or frequency of occurrence over an area or throughout a space or unit of time.
Differences	The degree or amount by which things differ in quantity or measure.
Relationship Outliers	A statistical observation that is markedly different in value from the others of the sample.
Dimension Reduction	The process of reducing the number of random variables under consideration by obtaining a set of principal variables.

3. METHOD

This research aim to identify and categorize the visual analytics representation based on the six types of relationship. There are 2 research activities involved.

Activity 1 has identified the context and problems through the literature review on the relationships in visual analytics such as correlation, distribution, comparison, differences, outliers and dimension reduction. The usage and functions of every charts, plots, tables and diagrams that is used to represent each relationship visualizations are determined to be categorized into which type of relationship later in activity 2. Besides, this study also identifying the relationship criteria that visual representation has to be achieved. From the previous studies, survey has been done to determine the level of end users understanding on relationship visualizations in organizations and the percentage of certified data analyst available in the current market as in 2018 which shows the need of this study for users to understand the relationship that shown in the visual analytics representations.

Activity 2 has collected all possible visual representation that exist in current market currently and divide it based on the types of relationship that the visualization is presenting. By doing so, we are able to determine all visual representation according to its own relationship group. We called this composition as Relationship Representation Catalogue. Currently, there is no proper collection or classification of graph according to its types of relationship. Hence, this catalogue will be able to tell us which visual representation can be use in representing which type of relationship. Furthermore, this study has identified the relationship representation criteria. There are six criteria has been determined which are: concretize relationship, determination of domain parameters, multivariate explanation, formulate cause and effects, confirm hypothesis, explicit the data to evidence. Thus, the classification of visual analytics representation is based on the types and criteria of relationship.

4. THE RELATIONSHIP CATALOGUES

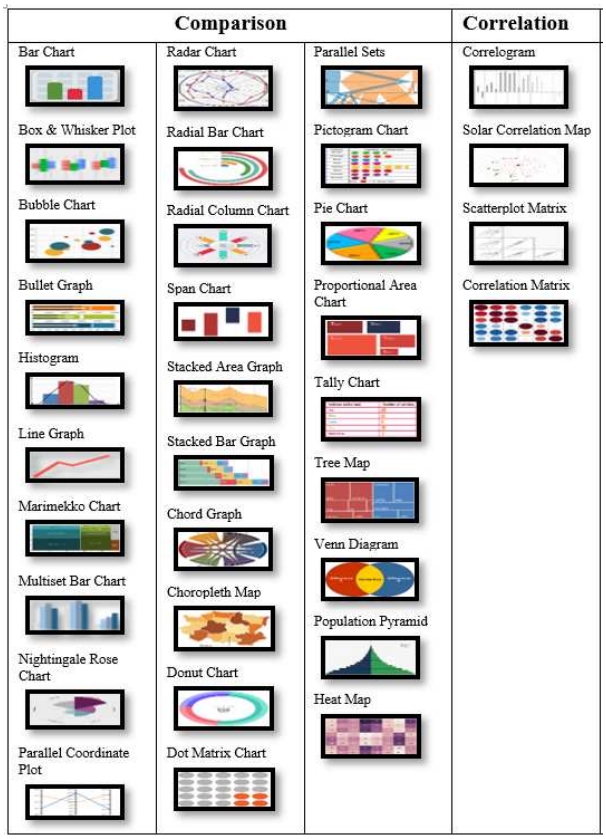
The finding from this research is the catalogue of visual analytics representation based on six types of visual analytics relationship. There is a total of 64 current visual representation (e.g. diagrams, charts, plots and tables) divided into six types of relationship in visual analytics. The selection and categorization of relationship are based on six relationship criteria as mentioned in Table 1. However, due to these characterizations, the research has eliminated 5 visual representations that are vague to represent the relationship elements. As for final categorization, only 59 visual representation are suitable to be analyze and divided into 6 relationship types as shown in Table 2. The paragraph below will describe each of the relationship representation in details.

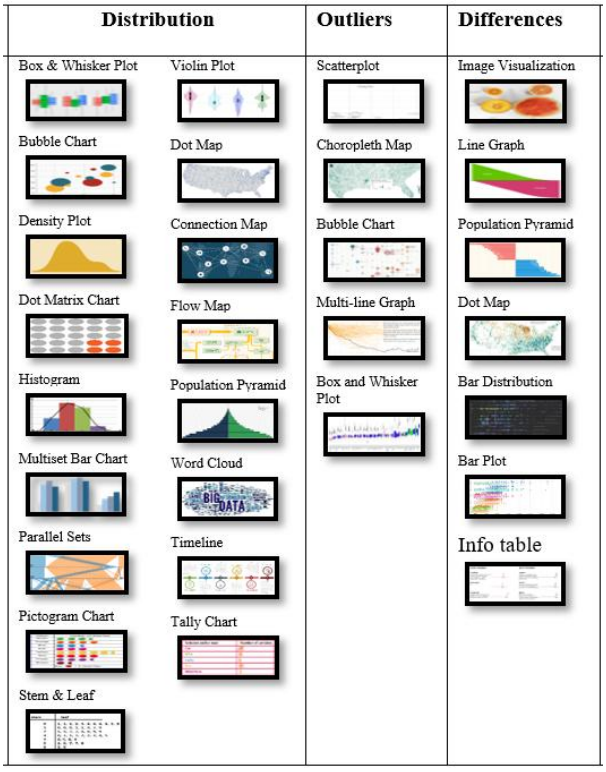
i) Correlation is meant to observe the relationship that exist between variables and able to show how strongly related they are. The research found several ways to visualize the correlation analysis such as correlogram, scatterplot matrix [10], parallel coordinated plot [11-12] and correlation matrix [13-14]. As one of the correlation analysis visualization - Correlogram is a plot based on the waveform-matching techniques that are often used in fundamental frequency extraction programs, but with no mechanism to select an actual fundamental frequency value [15]. Secondly is intercorrelation that highlight the importance of explanation between variables. Since the addition of many variables raises up the problem with dimensionality and requires large amount of data, then by choosing only one representative to represent a group of interrelated variables will help to visualize a meaningful intercorrelated data [16]. As an example, Solar Correlation Map use the analogy of solar system to show the intercorrelation between variables. By choosing one representative dimension as the center (sun) of solar system, then all other planets in each circle around the sun are representing the input variables whereas the moon are representing the input variables that are interrelated with the input variables (planets). The most popular visualization used by many in visualizing correlation relationship is correlation matrix. In correlation

matrix, each matrix cell shows the correlation value of a variable with the other variables. However, correlation matrix are not only used to visualize the correlation relationships, but it is also used in providing an overview of ranking features [13], support interactive filtering and clustering and also matrix reordering [14].

ii) Distribution relationship is to show the frequency of object or attributes in a collection of analysed data. From the literature analysis, the research found the concept of visualizing the distribution relationships is widely used in visual analytics. The techniques such as density plot, flow map, population pyramid, violin graph and histogram have been used widely to relate the distribution among analysed data. These techniques usually manipulate the elements of size, intensity, shape, hue and saturation of colours [17] to plot and show the distribution components. For quantitative data, the distribution analysis always constructed from numerical data, hence the elements of skewness, gaps, outliers, rounding or values are frequently occurred. As an example, density plot determined the highest concentration points of the data by using the peak value of the density plot. The usage of plots to distribute points along the numeric axis of the graph [18]. In the other hand, for highlighting the events, the research found the usage of timeline by visualizing the plots of events along the time axis where the event occur or on the range where the event last [19]. Timeline is used to represent the activities of an organization over the project time. Typically, timeline visualization is displayed in two or three dimensional which is useful when users want to have a quick view on the impact of upcoming tactical and operational activities. Apart from this, the research found the rise of popularity of visualizing distribution relationship using a violin plot. Violin plot is a combination of a box plot and density plot and it displaying the probability density and distribution of data [20]. The violin plot is eventually the mirrored, sideway density plot which have the same way of interpretation, the interquartile range represents by the box, the median can be seen as white dot in the middle and 95% confidence level is the whisker [21].

Table 2. Relationship Representation Catalogue





Compared to density plot that might create confusion in interpreting the data caused by the overlapping density plots, violin plot compares different datasets side by side and most importantly the shape of the violin is the values frequencies.

Finally, the Dot Map representation helps the user to easily detect the dispersion pattern. By manipulating the concept of spatial and ground, Dot Map will placed equally the sized points of the data over the geographical region [22]. Dot map consist of two types, one-to-one and one-to-many. One to one represents one single value over a dot whereas one to many represents several units over a dot. Dot maps are suitable in representing the distribution over the region and some patterns will be able to be point out when the points are in cluster on the map however, it is not great at retrieving the exact value. Compared to Choropleth mapping which assumes the random dot are the area the quantity are uniforms throughout the polygons, dot maps will eliminates the obvious graphic inaccuracies by placing the dots in a certain quantity of phenomenon where they are most likely to occur [23].

iii) Comparison relationship is intent to differentiate between several variables. This relationship will help the users to do an analysis tasks of estimation, measure, find the similarity and dissimilarity between analysis elements. Furthermore, it can help to examine the character or qualities of analysis in order to discover differences of resemblances among data. From the literature analysis, comparison relationship is the most popular and in demand of use. There are 31 types of visual representation have been developed and widely use in the market as listed in Table 1. Among the visual representation use to compare are charts (bar, radar, pictogram, pie, tally, bubble and radial), plot, graph (stacked area, bullet and line) and Venn Diagram. The comparison usually uses bar, line, size, area and radius of segment to represent the value of comparison. As an example, the manipulation of line in Line graph make the comparison more effective by visualizing the changes of data that occurs over time [24]. Besides, the patterns of the data are easily seen since it represents the data along the interval scale and it exists as single series line and multiple series line graphs. The customization of bar based on the users requirements and length of the bar will represent value that have been measured [25]. Bar chart is uses to show discrete and numerical comparison in the category by either vertical or horizontal bars. Different from histogram, bar chart did not show continuous development of interval, hence it will become problematic with the labelling when the value for a bar is large. Furthermore, bar element can be stacked together to display multiple series of graph that provides more insight compared to single series of graph. Other way of visualizing comparison relationship is using bullet graph. Bullet graph is a bar graph that has been improvised by Stephen Few, which it is a single metric that compares with other single metric, and visualized it in qualitative range [26]. Bullet graph objective is to produce the greatest amount of information with a limited space and reduced the confusion. It can be displayed either horizontal or vertical as long as it can be stacked to allow the comparison of the data at once. The research highlighted the usage of Nightingale Rose Chart as the combination of stacked column chart and radar chart [27].

Nightingale Rose Chart is drawn on a polar coordinate grid and radial chart. The equal segments of categories or interval are divided. In this chart, instead of the area of the segments, the radius of the segment represents the value. However, outer segments will stand out in the visualization because of their large area size, hence it will lead to an unreasonably represent increase in value. Different from Nightingale Rose chart, radar chart displays several quantitative variables over an axis like spokes on a wheel in circular. Starting from the centre, each variable is arranged radially maintaining the same scale between all axes. After all values are plotted, they are connected to form a polygon. However, it is only suitable to display data with limited variables since having too many variables will make it hard to be represents [28].

iv). Outliers relationship is to show an attributes or entities that situated away, isolated or detached from the system. In analytics, it is important to visualize the outliers as we want to show the factors or values that draws attention away from the averages. The literature analysis found 5 visual representations typically use to show the outliers which are scatterplot, choropleth map, bubble chart, multi-line graph and box and whisker plot. Commonly, there are five methods have been used to show the outliers; point of focus, breakout, scale adjustment, make an outlier as a reference point and providing context. While Point of focus lead the users to straight away identify the outliers, breakout method will use the concept of overview and details [29]. This method will display whole data and zoom to the outlier's value to be explained. The scale adjustment method has the same characteristics as box and whisker plot and finally, the method of taking the outliers as a reference point will make the data are more relatable to the outliers. Lastly, we can visualize the outliers by providing context. This method will display the data together with the outliers and reduce the ambiguous since we can distinguish from the data to the outliers. The challenge of outlier is when the values are large or small. When the outliers are too large or small, then it will hard to make it appear together with the normal data. Sometimes, the users might be wrongly count as a normal data.

v) Differences relationship is a straightforward visualization to distinguish the contrast, distinction, dissimilarity between analysis variables. This method aims for visualizations that allows quick comparison without having to do anything, just by looking at it, some information such as minimum, maximum, medium can be seen easily. Differences relationship also can be used as visual encodings that diverged to represent the relationships. Among visual representation use to distinguish the differences are image visualization (semiotics), line graph, population pyramid, info table, bar distribution and plot. Generally, the usage elements of geometry, colour, scale, coordinate, positive and negative sides will help to highlight the differences [30]. Among the methods use to show the differences are separate categories, side-by-side comparison, figure and ground, animation and sub-setting. By plotting several variables in each category, the differences will be obvious as it is side by side comparison. Instead of squeezing everything together, split things apart for side-by-side comparisons will make show the differences easily. Apart from this, relationship differences might be more effective by placing differences in the foreground and place the rest far in the background. By highlighting the differences, this method can simply subset some of the data to be visualize for audience. Lastly, animated contrast is the method used by the end users mostly, where it uses movement as the visual cue. It will show to the audience the intuitive view of how units shift to the other.

vi) Reduction relationship is the concept of trimming or limiting something in size, degree or amount. In reduction, the analysis techniques of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are widely used. PCA can be described as an unsupervised algorithm, since it "ignores" class labels and its goal is to find the directions that maximize the

variance in a dataset [31]. On the other hand, LDA is supervised and computes the directions that will represent the axes that that maximize the separation between multiple classes [32].

5. CONCLUSION AND FUTURE WORKS

As for now, there are many visual representations exist in visualizing relationship of an analytical data. Hence, there are still a lot of uncertainty of how the visual representation are providing enough support in analytical task and decision making. The analysis of relationship elements for each of the relationship type is critical to dictate the capacity of current visualization in representing the relationship of the analyzed data. In the future, this research will further analyze the relationship criteria and elements use in every visual representation based on each type of relationship catalogues. By reflecting to six relationship criteria and pre-attentive elements as mentioned in paragraph 2, this research will further identify how the relationship elements are presented, which visual analytics components used in highlighting those elements, and which relationship criteria that the elements are able to reflects through the relationship representation. These identifications will help the visual analytics community to have better understanding of the relationship and guide for improvement of visual analytics in term of representing the relationship in more complex information.

6. ACKNOWLEDGMENTS

The authors would like to thank Ministry of Higher Education (MOHE) and funded by Universiti Teknologi Malaysia (GUP Tier 2:Q.K130000.2656.17J23).

7. REFERENCES

[1] Shilpa M Patil. 1998. A Study on Visual Analytics. *Res. Gate* March (1998), 1–21.

[2] Gilbert Harman. 1984. Logic and reasoning. *Synthese* 60, 1 (July 1984), 107–127. DOI:https://doi.org/10.1007/BF00485621

[3] Pirolli Peter and Stuart Card. 2005. The Science of Analytical Reasoning. *Ieee* (2005), 33–68. DOI:https://doi.org/10.1057/ivs.2009.28

[4] Daniel Keim, Jörn Kohlhammer, Geoffrey Ellis, and Florian Mansmann. 2010. *Mastering the Information Age Solving Problems with Visual Analytics*. Retrieved May 28, 2019 from https://vismaster.eu/wp-content/uploads/2010/11/VisMaster-book-lowres.pdf

[5] Kristin A. Cook and James J. Thomas. 2005. Illuminating the Path: The Research and Development Agenda for Visual Analytics. Retrieved November 11, 2018 from https://www.osti.gov/biblio/912515

[6] Gilbert Harman. 1984. Logic and reasoning. *Synthese* 60, 1 (July 1984), 107–127. DOI:https://doi.org/10.1007/BF00485621

[7] Douglas N. Walton. 1990. What is Reasoning? What Is an Argument? *J. Philos.* 87, 8 (1990), 399. DOI:https://doi.org/10.2307/2026735

[8] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual Relationship Detection with Language Priors. . Springer, Cham, 852–869. DOI:https://doi.org/10.1007/978-3-319-46448-0_51

[9] Cook Kris Scholtz Jean, Burtner Russ. 2016. Visual Analytics 101. *CHI Ext. Abstr. Hum. Factors Comput. Syst.* (2016), 955–958. DOI:https://doi.org/10.1145/2851581.2856674

[10] William S. Cleveland and Marylyn E. McGill. 1988. Dynamic *graphics for statistics*. Wadsworth & Brooks/Cole Advanced Books & Software. Retrieved November 9, 2018 from https://dl.acm.org/citation.cfm?id=576082

[11] Robert M. Edsall. 2003. The parallel coordinate plot in action: Design and use for geographic visualization. *Comput. Stat. Data Anal.* 43, 4 (2003), 605–619. DOI:https://doi.org/10.1016/S0167-9473(02)00295-5

[12] Edward J. Wegman. 1990. Hyperdimensional data analysis using parallel coordinates. *J. Am. Stat. Assoc.* 85, 411(1990),664–675. DOI:https://doi.org/10.1080/01621459.1990.10474926

[13] Jinwook Seo and Ben Shneiderman. 1997. A rank-by-feature framework for interactive exploration of multidimensional data. *Inf. Vis.* 4, 2 (1997), 96–113. DOI:https://doi.org/10.1057/palgrave.ivs.9500091

[14] Harri Siirtola and Erkki Mäkinen. 1997. Constructing and reconstructing the reorderable matrix. *Inf. Vis.* 4, 1 (1997), 32–48. DOI:https://doi.org/10.1057/palgrave.ivs.9500086

[15] Svante Granqvist and Britta Hammarberg. 2003. The correlogram: A visual display of periodicity. *J. Acoust. Soc. Am.* 114, 5 (2003), 2934. DOI:https://doi.org/10.1121/1.1590972

[16] Zapf Stefan and Kraushaar Christopher. 2017. A new visualization to beautifully explore correlations - O'Reilly Media. Retrieved November 9, 2018 from https://www.oreilly.com/learning/a-new-visualization-to-beautifully-explore-correlations

[17] Sal Khan. 2012. Stem-and-leaf plots. Retrieved November 14, 2018 from https://www.khanacademy.org/math/ap-statistics/quantitative-data-ap/histograms-stem-leaf/v/u08-l1-t2-we3-stem-and-leaf-plots

[18] DatavizCatalogue. 2018. Density Plot - Learn about this chart and tools to create it. Retrieved November 10, 2018 from https://datavizcatalogue.com/methods/density_plot.html

[19] Catherine Plaisant, Brett Miiash, Anne Rose, Seth Widoff, and Ben Shneiderman. 1996. *LifeLines: Visualizing Personal Histories*. Retrieved November 14, 2018 from http://www.cs.umd.edu/projects/hcil

[20] DatavizCatalogue. 2018. Violin Plot - Learn about this chart and tools to create it. Retrieved November 14, 2018 from https://datavizcatalogue.com/methods/violin_plot.html

[21] Linh Ngo. 2018. 5 reasons you should use a violin graph - BioTuring's Blog. Retrieved November 14, 2018 from https://blog.bioturing.com/2018/05/16/5-reasons-you-should-use-a-violin-graph/

[22] DatavizCatalogue. 2018. Dot Map - Learn about this chart and tools to create it. Retrieved November 17, 2018 from https://datavizcatalogue.com/methods/dot_map.html

[23] A Jon Kimerling. 2008. Dot maps vs. choropleth maps with random dot area symbols. Retrieved November 17, 2018 from https://www.esri.com/arcgis-blog/products/product/mapping/dot-maps-vs-choropleth-maps-with-random-dot-area-symbols/

[24] Sweetspot. 2014. A focus on visualization: Line graphs. Retrieved Dec 2, 2018 from www.sweetspot.com/en/2014/05/13/focus-visualization-line-

graphs/

[25] Qualtrics Support. 2018. Bar Chart Visualization. Retrieved December 1, 2018 from www.qualtrics.com/support/survey-platform/reports-module/reports-section/reports-visualizations/chart-visualizations/bar-chart-visualization/

[26] Stephen Few. 2006. *Bullet Graph Design Specification*. Retrieved December 2, 2018 from https://www.perceptualedge.com/articles/misc/Bullet_Graph_Design_Spec.pdf

[27] AnyChart. 2018. Nightingale Rose Chart | Chartopedia. Retrieved December 2, 2018 from https://www.anychart.com/chartopedia/chart-type/nightingale-rose-chart/

[28] DatavizCatalogue. 2018. Radar Charts - Learn about this chart and tools to create it. Retrieved December 2, 2018 from https://datavizcatalogue.com/methods/radar_chart.html

[29] Ben Shneiderman. *The Eyes Have It: A Task by Data Type Taxonomy for* Information *Visualizations*. Retrieved June 14, 2019 from https://www.cs.umd.edu/~ben/papers/Shneiderman1996eyes.pdf

[30] Nathan Yau. 2018. Visualizing Differences | FlowingData. Retrieved December 3, 2018 from flowingdata.com/2018/04/17/visualizing-differences/

[31] R. A. Fisher. 1936. The Use of Multiple Measurements in Taxonomic Problems. *Ann. Eugen.* 7, 2 (September 1936), 179–188. DOI:https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

[32] Cheong Hee Park. 2008. Dimension reduction using least squares regression in multi-labeled text categorization. In *2008 8th IEEE International Conference on Computer and Information Technology*, 71–76. DOI:https://doi.org/10.1109/CIT.2008.4594652.