How would you visualise your data?

COMP5048/4448 Assignment 1

Albert Huang
530215563

Tutorial Activity 08
Katherine Mu

I. Task 1

A. What Do Those Data Represent/Describe?

The data selected in this assignment is "Data_Breaches_EN_V2_2004_2017_20180220.csv"[1]. The data describe details of major data breaches incidents between year 2014 and year 2017. The whole data set contains 270 incidents (rows) and 11 attributes (columns). In each row an incident of data breach is represented and information like name of organization, year, number of records lost from the breach, organization's sector... is stored in 11 columns. The three main data types are Nominal, Interval and Ratio.

B. How Were They Originally Collected?

The data used is derived from data "IIB Data Breaches - LATEST"[2], with improvements made like fixed inconsistent values and broken links, adjusted level of variables by Carlos E. Jimenez-Gomez[3]. The original data in "IIB Data Breaches - LATEST" is collected by David McCandless, Tom Evans, Paul Barton using sources from multiple media on "Information is Beautiful" website[4].

II. Task 2

A. Who Are the Consumers of Such Data?

The consumers of the data are likely professionals in cybersecurity, IT, risk management industries and official working in government security sector[3, 4].

B. Why Did They Need This Dataset?

They need this dataset to understand the key detail information about data breaches that have great impact. Attributes like reason, affected business sector, scale of impact would help them assess the incident in a clear and direct way.

C. How Are They Using This Dataset?

Through deeper analysis, patterns, trends, or key characteristics could be discovered for their benefit to conduct reports, counter measures, predictions. Visualization is a major usage of the dataset for them to present their findings.

III. Task 3

The data are typically depicted or conveyed to audience using bubble chart (examples shown at the end of this section). In this form of visualization, the main attributes are : name of the organization, name of sector impacted on, number of records lost, year, and methods of leak (reason). Number of records lost is ratio data, year is interval data and the rest is nominal data. Typically, all attributes' data are shown when user hover to a bubble as a highlight interaction. Specifically, the name of the organization are always shown for each bubble if there is enough space. The differentiation of sectors are done by assigning same color to bubbles with the same sector. Ratio data like number of records lost is shown by assigning sizes to bubbles accordingly. The visualization method can be justified using a visualization reference table by Łukasz Halik[5]. It is reasonable to represent selective perception, in this case, sectors, using color. And for quantitative perception like number of records lost, using sizes is a good way.

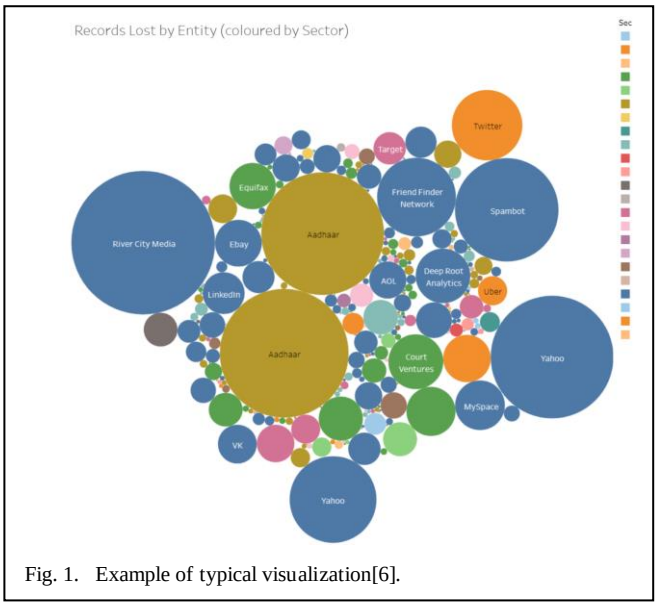


Fig. 1. Example of typical visualization[6].

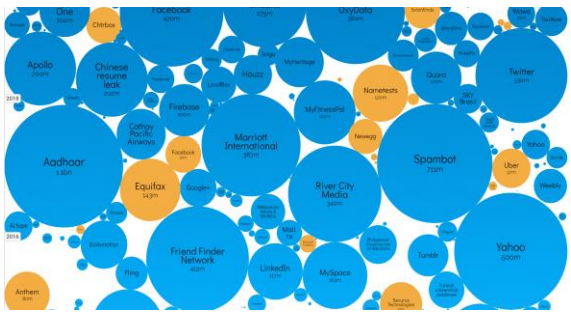


Fig. 2. Example of another typical visualization[4].

IV. Task 4

The questions that are typically asked and answered for the above visualization are: What quantities of records are compromised by organization, sector[3, 5]? What is the reason behind them[3, 5]?

V. Task 5

One typically mistakes people make is they use too much of the data[4] without proper filtering the data. There are visualizations using the original whole set of data (from 2004 to 2024) with more than 30 thousands of records, and this

cause an overwhelm perception of information when an user view the visualization.

Another mistake is poor selection of color[4, 6]. These visualizations is not color-blind friendly and there are uses of similar colors which make user hard to tell the difference.

Last but not the least, visualizations on the data are typically a lack of direct representation of attributes without user using interaction[4, 6]. Most visualization engage this matter by splitting relationships in multiple graphs, causing inefficient in representation[4, 6].

VI. Task 6

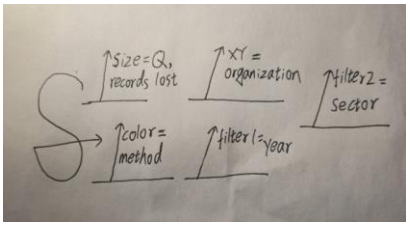
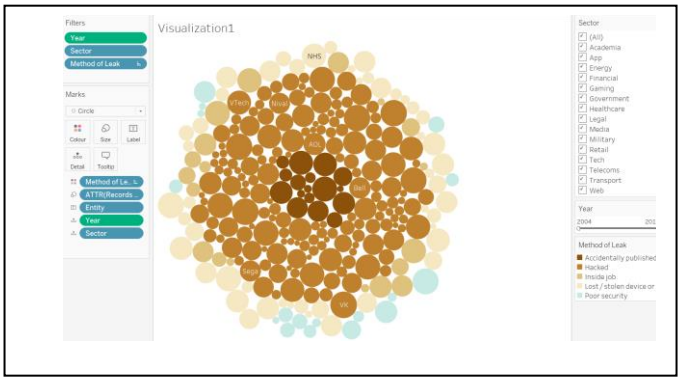


Fig. 3. Symbolic Representation

In this symbolic representation from Semiology of Graphics, the visualization should be in some kind of arrangement. It has quantitative variable Records Lost represented in size, Method of Leak represented in different colours, Entity (name of organization) represented as text, and two filters for year and sector for user to interact and represent these attributes by filtering.

VII. Task 7

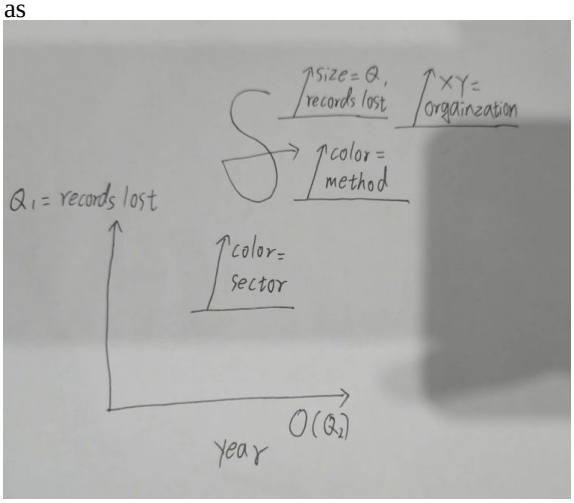


VIII. Task 8

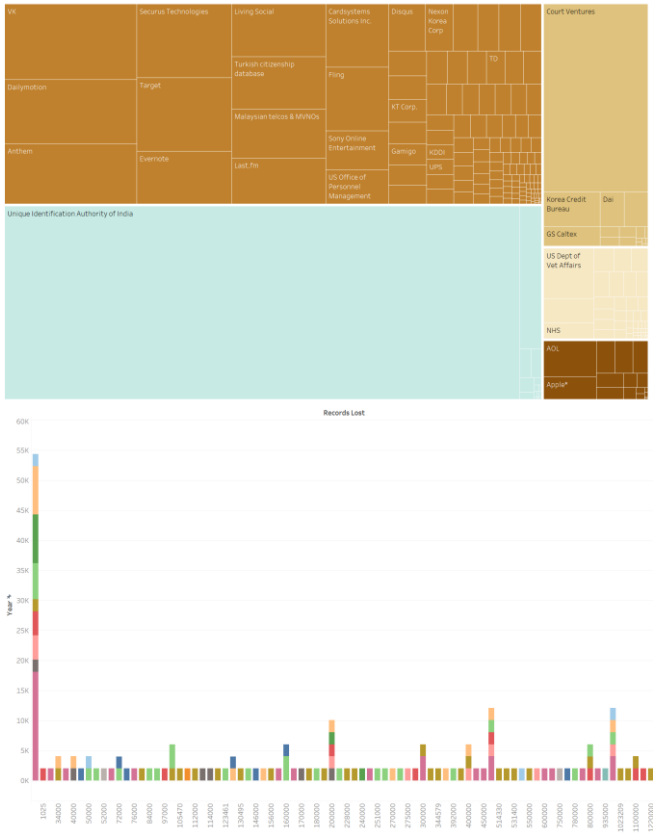
For a certain arrangement, combined with former typical visualization, it is reasonable to use bubble chart. Since Record Lost is a quantitative data and one major point of the visualization is to show this attribute in a direct way, assigning it with bubble size would be beneficial. To make colour recognizable and colour-blind friendly, Method of Leak is assigned to it because the total number of categories of Method is suitable for colour-blind friendly palette. And Method is a selective data. Entity and sector are assigned to text to fully express the information. Two filters are applied so the use can interact with the visualization to alter the way of presentation by year and sector according to their needs. These two attributes are represented in the form of interaction and comparison between actions.

IX. Task 9

In this case, I use two symbolic representations. Sector and year are derived to create another representation with Records Lost. Together, they are the same attribute with representations in task 6. And they share the same KGI as well as KPI.



X. Task 10



For graph one, it is the particial same as one in task 7 execpt using tree. For graph two, year and records lost are the two main axises. Records Lost as y axis since it is a quantitative data. And Sector is assigned with color for selective perception purpose.

References

[1] C. E. Jimenez-Gomez, "Data Breaches 2004-2017 (EN)," Kaggle, 2018. [Online]. Available: https://www.kaggle.com/code/xvivancos/data-breaches-tableau-visualization/input

[2] D. McCandless, T. Evans, and P. Barton, "Information is Beautiful," 2024. [Online]. Available: https://docs.google.com/spreadsheets/d/1i0oIJJMRG-7t1GT-mr4smaTTU7988yXVz8nPlwaJ8Xk/edit?gid=2#gid=2.

[3] C. E. Jiménez-Gómez, "Visualizing data breaches between 2004 and 2017," The Blog of Estratic, 2018. [Online]. Available: https://www.estratic.com/2018/02/09/visualizing-data-breaches-2004-2017-2/.

[4] D. McCandless, T. Evans, and P. Barton, "Information is Beautiful," 2024. [Online]. Available: https://informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/.

[5] Ł. Halik, "The analysis of visual variables for use in the cartographic design of point symbols for mobile Augmented Reality applications," Geodesy and Cartography, vol. 61, pp. 19-30, 2012. [Online]. Available: https://doi.org/10.2478/v10277-012-0019-4.

[6] Xavier, "Data Breaches Tableau Visualization," Kaggle, 2019. [Online]. Available: https://www.kaggle.com/code/xvivancos/data-breaches-tableau-visualization/notebook.