Swarm Plot: Data Redistribution in Non-Random Technique

Zainura Idrus
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Shah Alam, Malaysia
zainura@tmsk.uitm.edu.my

Fatin S. Rusli
Instantestore.com
Penang,
Malaysia
fatinsyahirahrusli98@gmail.com

Zanariah Idrus
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Kedah, Malaysia
zanaidrus@uitm.edu.my

Muhammad Aqil Mohd. Nazri
Intec International Education College,
Shah Alam, Malaysia.
aqilnazri9@gmail.com

Adel Al-zebari
Department of Information Technology
Duhok Polytechnic University
Duhok, Iraq
adel.ali@dpu.edu.krd

Noor Hasnita Abdul Talib
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Kedah, Malaysia
nhasnita@uitm.edu.my

2021 6th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE) | 978-1-6654-3402-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICRAIE52900.2021.9703994

***Abstract*— Data visualization is a process of converting data into visual forms with the purpose to extract hidden knowledge behind the data. One of the commonly used visualization is swarm plot which is suitable for small dataset. This is due to the fact that as data getting bigger, the data points will tend to overlap and be congested on the plot. Thus, it might lead to misrepresentation of data patterns and could result in false knowledge extraction. Thus, there is a need to have an effective swarm plot technique that can manage bigger data with no overlapping of data points. To achieve the objective, this research has formulated and enhances the process of identifying the right coordinate for each point on the plots. Overall, the system has successfully improved the redistribution of data points as such it can manage bigger data and at the same time avoid points from overlapping.**

Keywords — Swarm Plot, Point Redistribution, Big Data, Data Overlapping, Data Visualization

I. INTRODUCTION

Data visualization is a process of interpreting data by presenting it in a pictorial or graphical format. Data visualization helps people understand the significance of data by presenting it in a simple and easy-to-understand format to communicate the information clearly and effectively. With visualization, analysts can extract hidden knowledge behind the data through data patterns [1].

The swarm plot is one of the excellent visualization methods for better representation of data frequency distribution [2]. Swarm plot, also known as a bee swarm plot as it resembles bee swarming positions. Swarm plot is also very similar to a scatter plot where one of the axes is a qualitative variable [3]. Like scatter plot, swarm plot is a powerful data evaluation tool [4]

Swarm plot is usually plotted using point where one point represents one data on the graph. The point creates a swarm following the number of data used. Thus, the more the data, the larger the size of swarm forming [5]. This is where the problem occurs. The point's utilization space might overlap as points are packed to fit in the canvas [6]. Enormously, it results in smaller point as each point needs to be fitted into the canvas, thus, it is hard to decipher. Due to large data, swarm plot could be too big to occupy a canvas [7]. When these happen, they could cause distortion to the real frequency of data distribution and become a challenge during analysis phase.

Thus, there is a need to redistribute these points to assure there will be no overlapping and at the same time fitting the points into the canvas [5]

Random jittering technique is one of the techniques that has been widely used to avoid overlapping plotting. It randomly moves the data points away from one another. Nevertheless, the algorithm does not guarantee that over plotting will be avoided, and it often results in some points being moved carelessly [8]. Some of the jittering technique does not efficiently utilize the canvas space as there are excessive space between the points. While other generate too small points in a big canvas which could lead to perception issue as they look less dense. This resulted in low comprehensibility of data patterns on the plot.

Thus, this research aims to identify another redistribution technique and at the same time, improves the technique as such it could manage bigger data and avoid overlapping in swarm plot. To achieve the objectives, this paper is organized with Section I introducing swamp plot and it issues. Section II discusses some redistribution techniques as well as their strengths and weaknesses. The discussion continuous with Section III which describes the designing of the new improved redistribution technique. The results and discussions are in section IV where the focus is on the redistribution functionality testing. To sum up, Section V concludes this paper.

II. RELATED WORKS

A. Swarm Plot

The Swarm Plots, is also known as a bee swarm plot. It is similar to a scatter plot with one of the axes represent a categorical variable [9]. The swarm plot is being used mostly to compare variables and become the choice of people to ensure the plot do not look too statistical [10].

The plot has also been said to have the same advantages as the strip plot such that they plot every point in the dataset

978-1-6654-3402-7/21/$31.00 ©2021 IEEE

to give a good view of what is happening. A *bit random jitter* effects are often added in the strip plot so that the points will not overlap as much [11]. Unlike strip plot, swarm plot ensures the distance between plotted points are close to each other and do not overlap, thus giving the appearance of a swarm of bees [12]. Since the distribution of swarm plots do not overlap, it's easier to distinguish and visualize the frequency of data accurately.

B. Issues in Swarm Plot

The main objective of swarm plot is to plot points in a non-overlapped manner, thus, it is suitable for a relatively small number of data [12]. The data size plays a vital role in data distribution since a larger data generate a larger swamp. Sometime overlapping could not be prevented in larger data. Moreover, data with uneven distribution may also produce unpleasant swamp plot [13]. Fig. 1 shows an overplotting plot.

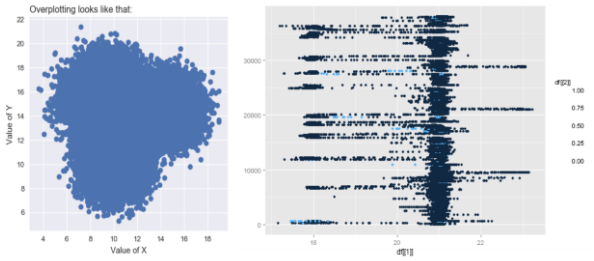

Fig.1: Overplotting
(Source: [14]*)*

The common scenario with swamp plot is that the plot space is limited to fit in a big amount of data. Thus. it is a challenge to formulate technique and calculation mechanism to distribution the data on the plot.

C. Techniques of Points Redistribution

This section discusses four redistribution techniques used to manage overlapping points. The four techniques are random jitter technique, stacking technique, non – random technique, and algorithmic jitter technique.

- Random Jitter Technique

Random jitter technique is performed by adding random noise to data to prevent points from being overplotted on top of each other [15]. This technique is usually used to highlight data distribution and data frequency. However, it is suitable if the exact location of the point is not important for analysis. This is because the main concern of the technique is on preserving the data frequency within a specific region [16]. Fig. 2 shows plots without the random technique while Fig. 3 is after the technique implementation [17].

Since jittering is focusing on ensuring points are not overlapping, space between points is rarely taken into consideration [14]. Moreover, the point size is also not in the concerned. Bigger point will lead to overlapping plot and too small point leads to perception issue since plenty of space between points. Thus, it is often perceived as less frequency.

- Stacking technique

For stacking technique, each point is with equal values and they are stacked on top of each other vertically [18]. This technique can support a great amount of data [19]. An example in Fig. 4 shows a dot plot for a bimodal dataset that used the stacking technique.

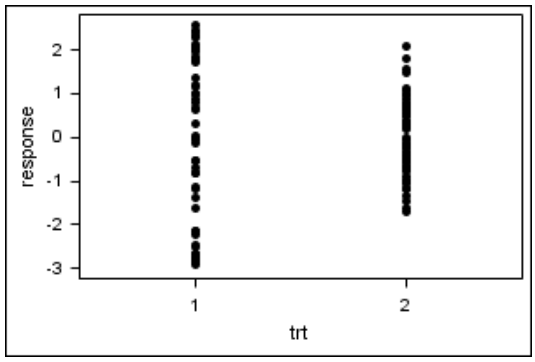

Fig.2: Strip plot without random jitter technique
(Source: [17]*)*

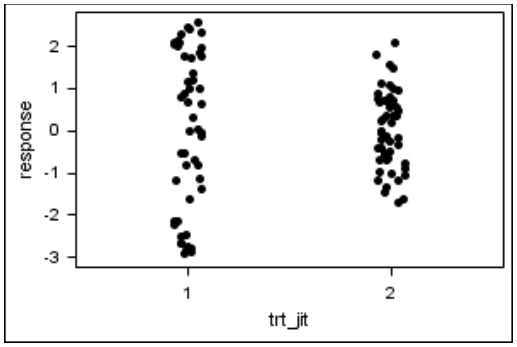

Fig 3: Strip plot with random jitter technique
(Source: [17]*)*

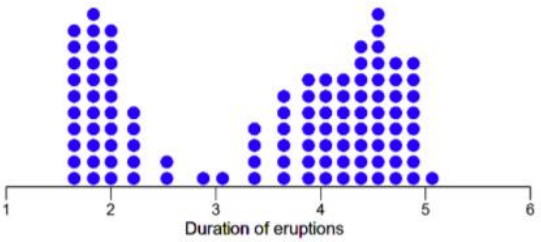

Fig. 4: Stacking
(Source:[19]*)*

- Non – random technique

Non - random technique calculate the distance between two points to avoid overlapping. This computational algorithm is based on distance. However, it is executed after jittering strip plot has been executed. The distance algorithm is executed repeatedly until every data point has found its location and no overlapping occur [20]. Fig. 5 and Fig. 6 show jittered strip plots before and after respectably the distance algorithm has been applied.

- Algorithmic Jitter Technique

Algorithmic jitter technique plots data points with a minimum overlap to its neighbors' points. It calculates the percentage of overlapping points which have the same value. The higher the percentage of overlapping means that there are more points that exist with the same value. This will keep repeating until the maximum overlap is reached. During

plotting, the same principles are also applied to a higher and a lower data value. This makes the data points not randomly jittered but iteratively placed within outlines of the functional shape of the data [15]. Fig. 7 illustrates the approach. Algorithmic jitter technique takes less space as compare to random jitter technique.

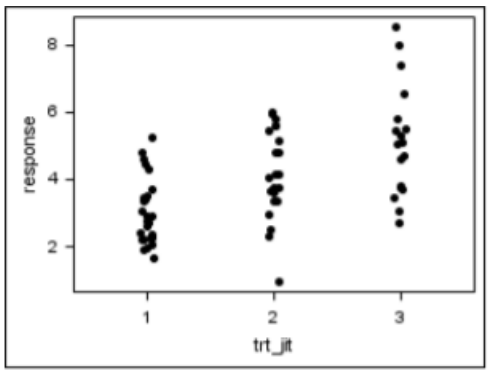

Fig. 5: Jittered Strip Plot
(Source: [20]*)*

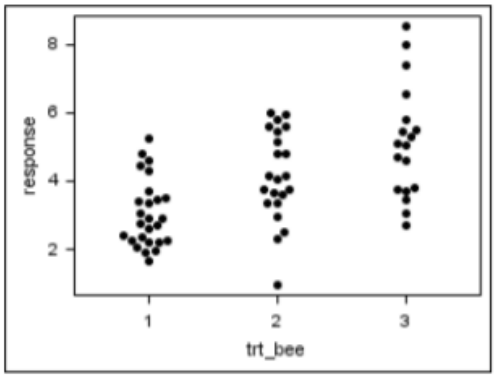

Fig.6: Beeswarm Plot
(Source: [20]*)*

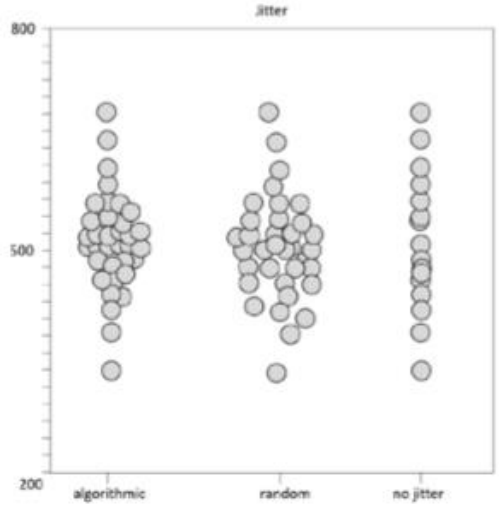

Fig. 7: A comparison of the algorithmic jitter function and random jitter, opposed to the plot without jitter
(Source: [15]*)*

D. Comparison of Redistribution Techniques

This section compares the strengths and weaknesses of the redistribution technique in an attempt to adopt the best technique for better result. The comparison is in Table 1.

Table 1: Advantages and disadvantages of redistribution techniques

Redistribu-tion Techniques	Advantages	Disadvantages
Random Jitter Techniques	• reduces overplotting of ordinal data • better visualize the density of the data and the relationship between variables. • displays clusters in the data.	• no unique way to jitter. • distribution of random component is not always clear. • not suitable if the exact location of a point is important to the analysis.
Stacking Technique	• support 2D and 3D visualization. • not limited to points and lines. • no overplotting.	• not scale well to a very large data set. • small gaps in spaces with a high dot density are smoothed out because of the constant dot size, and therefore become unperceivable.
Non - Random Technique	• many points will fit vertically and horizontal. • no overplotting.	• time-consuming as the algorithm continues until every data point has a location without overlapping. • computationally expensive as it needs a considerable number of resources like time, processing power, memory, etc.
Algorithmic Jittering Technique	• the width of the jitter is controlled by the density distribution of data. Data points were iteratively placed within the outlines of the functional shape of the data. • reduces overplotting	• finding the optimal setting for a specific set of data and use cases is difficult. • computationally expensive as it needs a considerable number of resources like time, processing power, memory, etc.

(Source: [15], [19]–[22]*)*

Table 1 shows the advantages and disadvantages of each technique. Based on the observation and simulation, the random jitter techniques and algorithmic jitter techniques could reduce the over plotting but not completely get rid of it as a larger number of data can still create overplotting. The stacking technique ensures that there will be no overplotting but with large data it will make space between the point blurred as the points will become much smaller. The non-random technique ensures that there is no overplotting as the

algorithm calculate the distance between each point to suit the size of a canvas.

Therefore, a non-random technique is the most suitable technique to be used on the swarm plot with the intention to solve overplotted issue. It can reduce overlapping better as compare to random jitter, algorithmic jitter while stacking could not support large dataset.

III. ALGORITHM DESIGN

The swarm plot used in this research is a 2-dimensional graph. There are eight steps in developing a swarm plot as depicted in Fig. 8. The first step begins with getting data. Then SVG are created for the graph. The SVG is the medium used to display the graph to the user. In the third step, the minimum value and maximum value were chosen from the data. The minimum and maximum values are used in the fourth step, which is creating an axis for the graph. Next, set the radius size that seems suitable for the points, based on the data size. In the next step, data points are calculated to generate their coordinate on graph. The points are then mapped based on the predefined coordinate. Finally, color is added to increase visibility of data distribution.

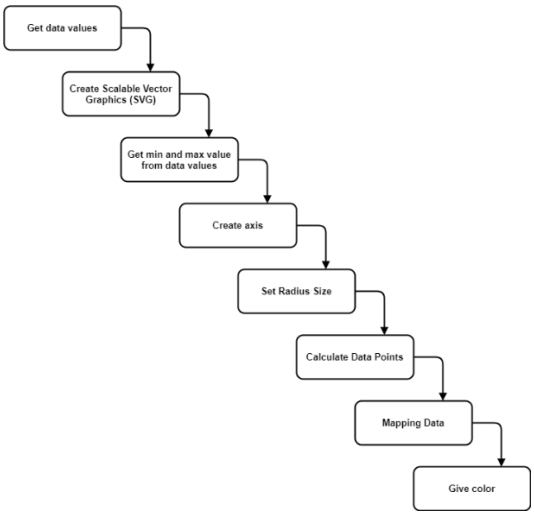

Fig.8: Steps of developing Swarm Plot

Two important steps in creating swamp graphs are the formation of point size and the location of the points. This research has improvised these two steps to address the issue of points overlapping in bigger data.

A. Formation of Point Size

For point size, in most cases the point radius is manually fixed based on try and error process during development stage. However, with such method, it could end up with overfit or under fit issue. If data are big, overlapping could occur. On the other hand, if data are less, there will be too many spaces between points.

Thus, there is a need to formulate a dynamic radius size. The dynamic point size is calculated based on the formula

$$10 - Math.floor(Math.log(data.length)) \quad (1)$$

The point radius is formulated based on the quantity of data. With that, point radius varies depending on data used.

B. Location Determination

For point location, the x coordinate uses the values of the data while the y coordinate is the same for every data point. Y coordinate are generated by dividing the graph height by two to get the center of the graph. From the center, a non-random technique distribution is applied in order to avoid overlapping of points. With a non-random technique, a suitable distance between two points needs to be calculated where distance formula has been applied as shown in Fig. 9.

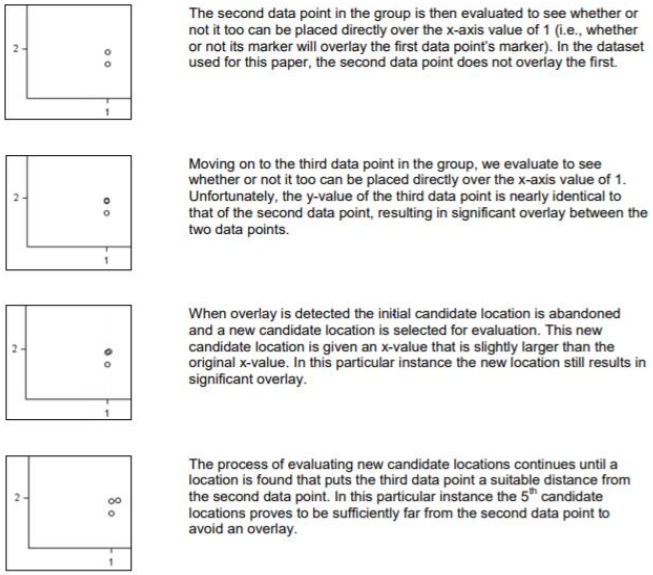

Fig. 1 Distance formular in a non-random technique (Source: [20]*)*

Overlapping nodes are resolved through iterative relaxation where anticipated positions formular

$$(*x* + *vx*, *y* + *vy*) \quad (2)$$

Where *v* is velocity and *xy* are the coordinate of the swamp plot. The point's velocity is then modified to push the points out of each overlapping points.

The process has to be repeated for a certain number of looping. However, the number of looping is decided manually, thus, it is fix. With this practice, the technique could not be generalized, thus it is suitable to a limited data size. In most cases, partial overlapping could not be avoided. This is true for bigger dataset.

With that in mind, a generalized number of repetitions have been formularized. Where the number of repetitions is calculated based on the size of the data.

$$n = data.size \quad (3)$$

where *n* is the number of repetition and size is the number of data need to be plotted.

IV. RESULTS AND DISCUSSION

This section elaborates the testing stage in order to ensure that the new formulation of data allocation works well in accordance with the demands and also to prevent any conditions that could jeopardize the system. The test findings will be discussed and analyzed.

A. Small Sample

The data used is called Medical Cost Personal Datasets. To generate a small dataset, a portion of the data have been selected which comprise of 315 data.

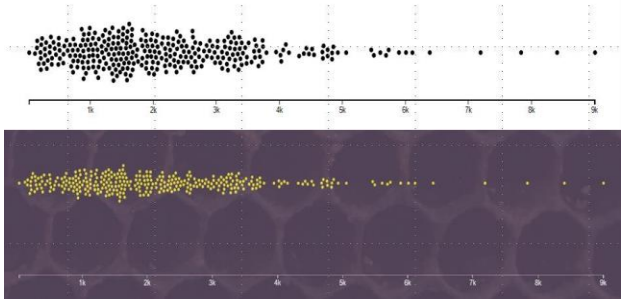

Fig. 10: Before and after the improvise technique for small dataset

Fig. 10 shows the graphs of before and after improvised non-random technique. Both swarm plots are nicely plotted where there is no overlapping occur. In term of point size, the improvised technique produces smaller points. Similarly, with the improvised technique, the space between points has been reduce. Each point become closer to one another. Both scenarios are acceptable as they do not distort the data pattern. The data patterns are the true knowledge hidden behind the data. It can be concluded that both methods successfully plot the data as there is no overlapping or disruption to the data frequency.

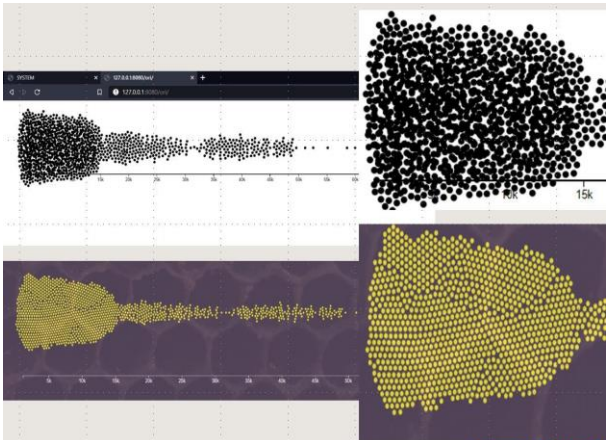

Fig. 11: Before and after the improvise technique for bigger dataset

B. Bigger Sampling

Next, testing is conducted on bigger dataset as depicted on Fig. 11. A total of 4000 dataset of Medical Cost Personal has been used. In the previous technique, the number of data seems larger because of the crowded, bigger point size and overlapping points. This could lead to misinterpreted data frequency. On the other hand, the new improved method, plots all point successfully without overlapping as compared to the previous method. The new improvised technique has successfully reduced the point size as well as spaces between points. Moreover, the redistribution of point has also smoothly reallocated and the data patterns are preserved.

It can be concluded that the enhancements of the technique have successful manage bigger dataset as compared to the previous version.

C. Others Comparison

The testing continues with COVID-19 Community Mobility Reports Datasets. It is built up from 3024 dataset. Overall, the graphs are not very efficient, and a few problems can be seen as compared to the enhanced version.

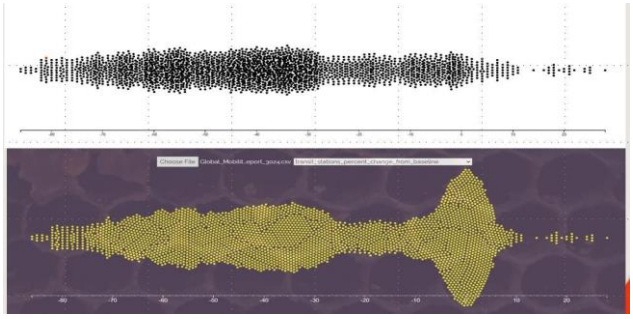

Fig. 12: Before and after the improvise technique for bigger dataset

There is an overlap happening in the original graph which disturbs the frequency. In addition, the shape of the swarm plot is totally different, as can be seen in Fig. 12. This is because of the crowded and overlap points, making the graphs overall poorly constructed. This shows the disturbance in the data frequency. However, with the new improvised method, the true frequency could be revealed.

D. Comparison Analysis

The Table 2 shows an analysis of each dataset's swarm plot. Based on the table, it can be concluded that all three datasets worked with the improvised version of swarm plot.

Table 2. Comparison Analysis

	Original Swarm Plot Method	Enhancement Swarm Plot Method
Medical Cost Personal Dataset (315 data)	No overlap points happen in the plot	No overlap points happen in the plot
Medical Cost Personal Dataset (4000 data)	There are overlap points happen in the plot	No overlap points happen in the plot
COVID-19 Community Mobility Reports Datasets (3024 data))	There are overlap points happen in the plot	No overlap points happen in the plot

There is no overlapping occur on all three datasets when they are fed to the improvised version of swarm plot . The original algorithm is seen to work with small dataset only. The data size also plays an essential role in ensuring no overlap points in the swarm plot. The larger the data, the more overlap points in the plot.

V. FUTURE WORK

The new improvement algorithm could only accept numeric values. String values are not accepted and are filtered out by the system. This is because the currently used algorithm only calculates numeric data. Recommendations

for enhancement is for the swarm plot to visualize string data by groups.

VI. CONCLUSION

The objective of this research is to identify a distribution technique which could be adopted to swarm plot. Various distribution techniques such as random jitter technique, stacking techniques, non-random technique and algorithmic jitter technique have been studied. Out of those techniques listed, non-random technique is the one that was chosen to be implemented in the swarm plot. This research has improvised the non-random technique for efficient use of plotting space as such more data points could fit into the plot without overlapping. Finally, our testing has proven that the new improvised redistribution technique can accommodate bigger data without overlapping.

ACKNOWLEDGEMENT

The authors would like to thank Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Selangor, Malaysia for facilities support.

REFERENCES

[1] Z. Idrus, Z. Idrus, S. Z. Zainal Abidin, N. Omar, and N. S. A. Mohamat Sofee, "Erratum to: Geovisualization of Nonresident Students' Tabulation Using Line Clustering," in *Regional Conference on Science, Technology and Social Sciences (RCSTSS 2016)*, Singapore: Springer Singapore, 2018. doi: 10.1007/978-981-13-0074-5_105.

[2] M. Burch *et al.*, "The Power of Linked Eye Movement Data Visualizations," May 2021. doi: 10.1145/3448017.3457377.

[3] A. Yu, C. Chung, and A. Yim, *Matplotlib 2.x By Example*. Packt Publishing, 2017.

[4] D. Park, S. M. Drucker, R. Fernandez, and N. Elmqvist, "ATOM: A Grammar for Unit Visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 12, Dec. 2018, doi: 10.1109/TVCG.2017.2785807.

[5] Brock. Tim, "Too Big Data: Coping with Overplotting," *infragistics*, Apr. 20, 2015.

[6] F. Qiao, "Large Scale Visualizations and Mapping with Datashader," *Toward Dta Science*, Dec. 25, 2018.

[7] N. Yau, "How to Make Beeswarm Plots in R to Show Distributions," *Flowingdata*, 2016.

[8] W. A. W. R. I. Z. I. Z. and I. S. N. Ramli, "Data visualization: Toward visualization awareness in overlapped dots management techniques," *Journal of Theoretical and Applied Information Technology*, vol. 97, no. 12, p. . 3390-3401, 2019.

[9] A. Cole, "Seaborn Visualizations Tutorial. A walkthrough of many Seaborn tools… | by Andrew Cole | Towards Data Science," *Towards Data Science*, Mar. 23, 2020. https://towardsdatascience.com/seaborn-visualizations-tutorial-c390bf1d43ce (accessed Jul. 26, 2020).

[10] I. Dillingham and A. McCarthy, "Creating a Buzz around Corporate Reputation with Beeswarm Plots," 2016.

[11] A. Gude, "Visualizing Multiple Data Distributions | Alex Gude," *alexgude.com*, Apr. 24, 2017. https://alexgude.com/blog/distribution-plots/ (accessed Jul. 26, 2020).

[12] M. Döring, "Box Plot Alternatives: Beeswarm and Violin Plots - datascienceblog.net: R for Data Science," *R for Data Science: datascienceblog.net*, Nov. 04, 2018. https://www.datascienceblog.net/post/data-visualization/boxplot_alternatives/ (accessed Jul. 26, 2020).

[13] A. Eklund, "The Bee Swarm plot, an Alternative to Stripchart (Internet)," *R Foundation for Statistical Computing, Vienna*, 2003.

[14] F. Qiao, "Large Scale Visualizations and Mapping with Datashader | by Finn Qiao | Towards Data Science," *Towards Data Science*, Dec. 25, 2018. https://towardsdatascience.com/large-scale-visualizations-and-mapping-with-datashader-d465f5c47fb5 (accessed Jul. 26, 2020).

[15] M. D. Kickmeier-rust, "Using Jitter and Sampling Techniques to Improve the Comprehensibility of Scatter Plots : A Practical Example," in *Proceedings of the Learning Analytics and Knowledge Conference (LAK'19), March 4-8, 2019*, 2019, pp. 1–5.

[16] W. A. W. Ramli, Z. Idrus, Z. Idrus, and S. N. Ismail, "Data visualization: Toward visualization awareness in overlapped dots management techniques," *Journal of Theoretical and Applied Information Technology*, vol. 97, no. 12, pp. 3390–3401, 2019.

[17] Rho, "I Swarm, You Swarm, We All Swarm for Beeswarm (Plots)," *rhoworld.com*, Nov. 21, 2014. https://www.rhoworld.com/i-swarm-you-swarm-we-all-swarm-for-beeswarm-plots-0/ (accessed Jul. 26, 2020).

[18] I. E. Frank and R. Todeschini, *The Data Analysis Handbook*. Elsevier Science, 1994.

[19] T. N. Dang, L. Wilkinson, and A. Anand, "Stacking graphic elements to avoid over-plotting," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1044–1052, 2010, doi: 10.1109/TVCG.2010.197.

[20] S. Rosanbalm, "A Strip Plot Gets Jittered into a Beeswarm," *rhoworld.com*, pp. 1–12, 2014.

[21] L. Akrofi, "When to Use and Not Use a Jitter Plot on Tableau Visualization," *The Data Crunch*, Aug. 24, 2019. https://datacrunchcorp.com/jitter-plot/ (accessed Jul. 26, 2020).

[22] R. Wicklin, "To jitter or not to jitter: That is the question - The DO Loop," *blogs.sas.com*, Jul. 06, 2011. https://blogs.sas.com/content/iml/2011/07/06/to-jitter-or-not-to-jitter-that-is-the-question.html (accessed Jul. 26, 2020).