Data Visualization: Trends and Patterns

Abstract

Data visualization is a broad term encompassing various techniques aimed at enhancing people's comprehension of data by presenting it visually. It transforms quantitative information into graphical representations, making it easier for the human mind to identify patterns, trends, and correlations that might otherwise remain hidden within text-based data. Although data visualizations frequently take the form of familiar charts and graphs, they play a prevalent role in our daily lives. Moreover, they have the potential to reveal previously undiscovered insights and trends. The art of crafting effective data visualizations combines elements of communication, data science, and design, offering valuable and intuitive insights into complex datasets. In this article, we will delve into the world of data visualization, exploring its significance, tools, and applications.

Authors

Ashish Gupta
Research Scholar
Department of Computer Science & Engineering
RNT University, Bhopal, India
guptaashishnitm@gmail.com

Dr. Sanjeev Kumar Gupta
Dean, Engineering
RNT University, Bhopal, India
sanjeevgupta73@yahoo.com

Dr. Pritaj Yadav
Associate Professor
Department of Computer Science & Engineering
RNT University, Bhopal, India
yadavpritaj@gmail.com

Dr. Deepak Gupta
Associate Professor
Department of Computer Science & Engineering
ITM, Gwalior, India
deepak.gupta@itmgoi.in

I. Introduction

Data pattern recognition plays a crucial role in various industries, particularly in pharmaceuticals and healthcare. Although there are software tools available to automate this process, and machine learning can handle complex data, the manual review of data remains essential, even for simpler aspects of these sectors. This includes the evaluation of metrics like batch record, microbial counts, and categorization of divergence, among other factors. Data visualization serves as a valuable and universally understandable means for pattern analysis (1).

This article introduces several approaches for gaining perception from data utilizing straightforward imaged Tools. While these imaged. tools are uncomplicated, they excel in conveying essential points with clearness (2) .

II. What is a Data Pattern

We attempt to break down an issue one essential aspect is the search for discernible patterns within the generated data. Patterns are essentially similarities or shared characteristics that meet specific criteria. Complicated shapes recognition is a basic concept in information technology and can also be applied to set data using software tools like MS Excel.

Data pattern within a dataset is essentially a sequence of data points that repeat in a understand manner. This recognition may be based on the historical data being analyzed or on data that exhibits identical characteristics. The easy patterns often involve numerical values that exhibit either upward or downward trends. These patterns become more evident when the numerical data is visually presented in graphs or tables. Patterns can also be identified through basic statistical analysis, such as searching for correlations among two sets of no.

Two commonly encountered types of data patterns are those associated with time (e.g., seen in trend charts) and those linked to causality (e.g., observed in regression analysis). Time sequence models consider that the direction a chart takes is primarily associated to its own historical patterns, while models assess the relationship among other influencing factors as well as the data under consideration.

III. Understanding Data Collection Methods

It is crucial to comprehend how data was collected and its relevance before delving into pattern analysis. Data often falls into one of two categories:

Cross-Sectional Data: These are apprehensions gathered at a specific point in time, such as a sequence of tests conducted on a single in-process sample.

Time Sequence Data: These are gathered over successive time intervals. For instance, it could involve a sequence of in-process samples tested at various points in time in relation to a particular test.

The manner in which data is collected dictates the types of patterns that can be explored. In the case of time-dependent data, there are four overarching pattern types: trending, seasonal, horizontal and cyclical (3). Conversely, with cross-sectional data, the emphasis is further on extracting information and identifying patterns within individual events.

IV. Time and Trends

When a substantial number of data points are available, and these data are collected over a relevant time frame, it becomes possible to identify a trend. A trend represents the long-term component that signifies either enhancement or decline within the time sequence over an extended duration. Line chart is particularly well-suited for visualizing continua data, as they connect numerous data points that all pertain to the same category.

In the case of these chart types, data points may exhibit slight variations, but on the whole, the data exhibits a consistent direction. For instance, Figure 1 illustrates the increase in microbial counts for the same sample calculated over a period of time 1.

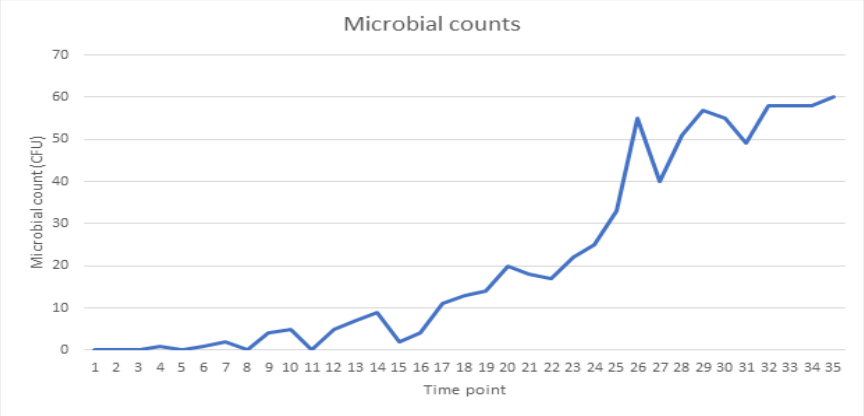

Figure 1: Data were Analysed Over Time for Microbial Counts.

Alternatively, when examining pH readings collected from similar process point across consecutive batches, the data reveals a consistent decline across multiple time point, as depicted in Fig. 2.

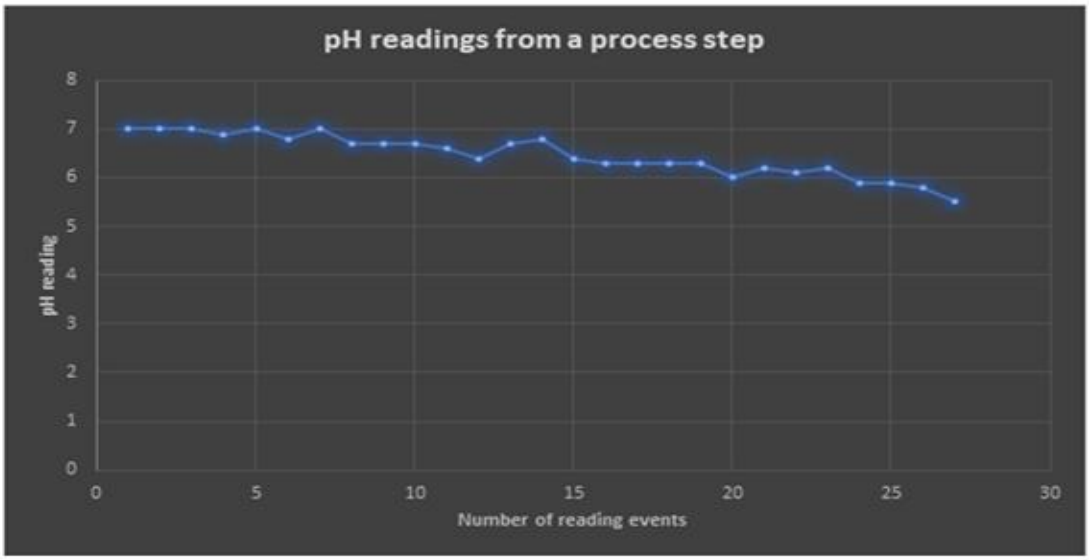

Figure 2: The Data Analysis for Ph Readings over Time

Additionally, time-situated data can be analyzed to identify its cyclical characteristics. The cyclical element represents the wavelike fluctuations around the underlying trend. To illustrate this, let's revisit the microbial count data, that exhibited improvement over time subsequently corrective actions. In this case, a clear cycle which can be observed, as demonstrated in Figure 3.

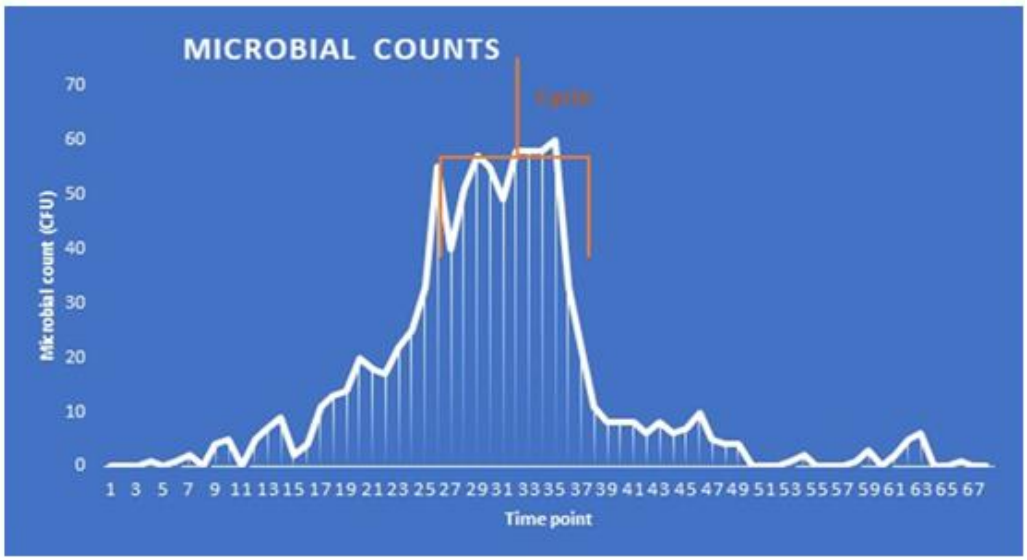

Figure 3: A Data Cycle is Focused on (In Regard to Microbial Counts)

Cycles can be associated with various events, including intervals within the broader time frame depicted in the graph, such as a month or quarter, or year, or in connection with controlled changes. Additionally, cycles can manifest as long-wave shapes, and these occasionally exhibit repetition. For instance, consider microbial counts from a water system, as illustrated in Fig. 4.

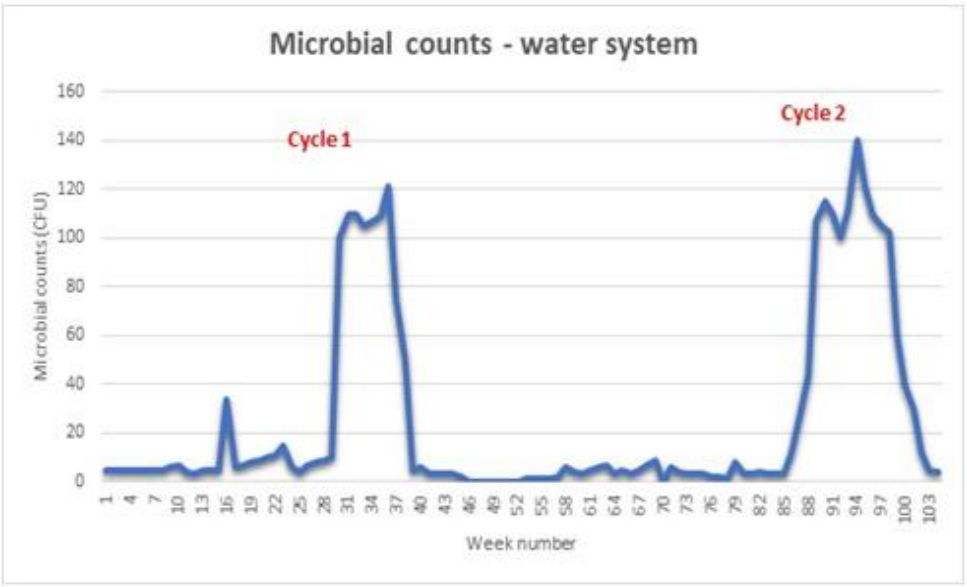

Figure 4: Example of a Seasonal Cycle for Microbial Counts that is repeated Over Time

In this case, two distinct cycles emerge over a span of 2 years, coinciding at roughly the similar time each year. This observation might indicate that there is a specific period (such as summer) during which microbial counts tend to rise. In a hypothetical scenario, these count increases could be tied to a production shutdown for preservation reasons. Often, such cyclical patterns exhibit regularity, although their durations can vary.

These cyclical patterns can transition into seasonal elements, characterized by a repetitive pattern that recurs year after year. When data gathered over time exhibit fluctuations around a constant level, a horizontal pattern is present. Such a series is considered to have a stationary mean. For instance, if monthly yields for an active pharmaceutical ingredient remain relatively constant without a consistent increase or decrease over a period, they will be classified as having a horizontal pattern.

V. Explaining Data Patterns

When evaluating gathered data, it's valuable to provide descriptors that help characterize the data:

1. Are the data random, where successive of a time series lack any discernible relationship?
2. Do the data exhibit a trend, indicating they are nonstationary?
3. Are the data stationary or horizontal?
4. Do the data display seasonality?

Using these descriptors, a series that fluctuates around a consistent level without showing enhancement or decline over time may be termed "stationary." Consequently, a stationary time series maintains constant statistical properties, such as mean and variance, as time progresses. Conversely, a series that includes a trend can be categorized as "non-stationary."

VI. Data Distribution

Another perspective on data involves examining its distribution. Graphic displays are a valuable tool for visualizing patterns within data. This visual analysis can be applied around a work to make informed decisions or adjustments regarding design and study variables while maintaining experimental control and achieving enhanced outputs. Additionally, it can aid in assessing data for normality or other characteristics before selecting the properly statistical analysis tool.

Patterns within data distribution are typically described in forms of four key aspects: center, shape, and any unusual features (4). When graphed, the center of a distribution, where the central data is concentrated, corresponds to the median of the distribution (as seen in Figure 5). This median represents the point in a graphical presentation where roughly half of the observations fall on either side. In the chart below,

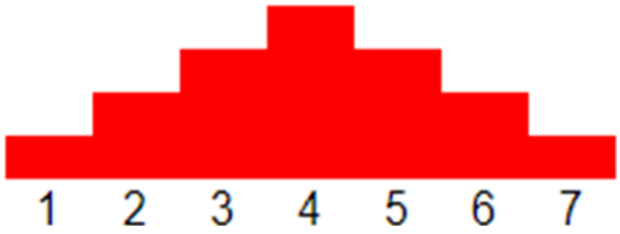

Figure 5: Data Demonstrating a Centralised Distribution

Similarly, a uniform distribution occurs when observations within a dataset are evenly distributed across the entire range of the distribution, as illustrated in Fig. 6. In a homogeneous distribution, there are no distinct peaks or concentrations of data.

Figure 6: Data Demonstrating a Consistent Distribution

The spread of a distribution pertains to the extent of data variability. When observations encompass a broad range, the spread is more extensive, as demonstrated in Fig. 8. Conversely, if observations cluster closely over a single value, the spread is narrower, as evident in Fig. 7.

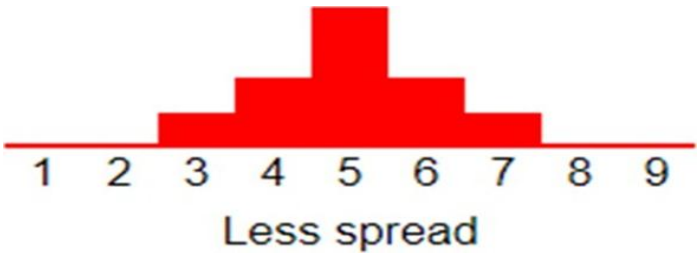

Figure 7: Data that have a Small Spread

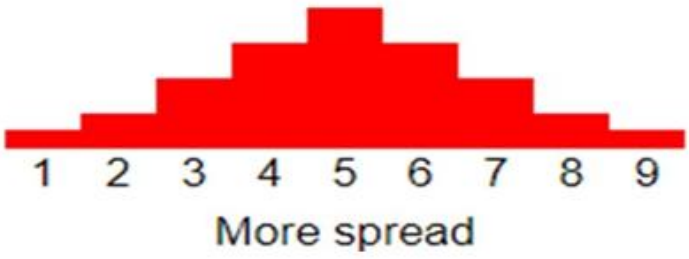

Figure 8: Data with a Comparatively Wide Spread

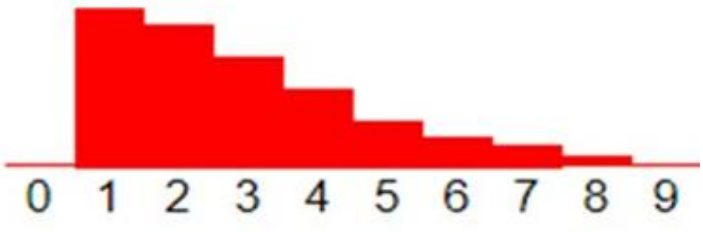

Figure 9: Data that are Skewed to the Right, As if this were Frequently the Case for Microbial Data

Skewness: When graphically represented, certain distributions exhibit a notable imbalance, with a considerably greater number of observations on one side of the graph compared to the another. Distributions with fewer observations on the right side are described as having a right skew, while those with fewer observations on the left side (toward lower values) are characterized as having a left skew. Microbiological data, for instance, often displays right skewness, as demonstrated in Figure 9.

Outliers: On occasion, distributions include extreme values that significantly deviate from the rest of the observations. Outliers are the terms used to describe these extreme values. According to general guidelines, an extreme value is typically regarded as an outlier if it is at least 1.5 interquartile ranges either above or below the first quartile (Q1) or third quartile (Q3). Figure 10 shows an illustration of such an anomaly.

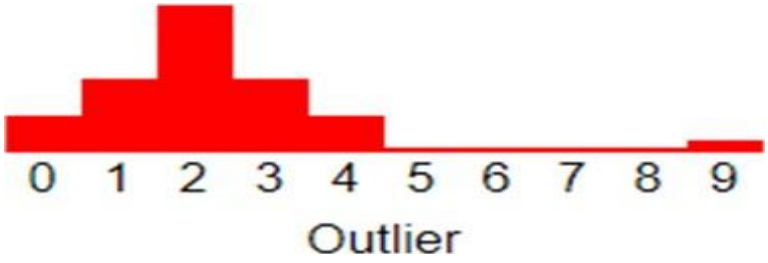

Figure 10: Graph of the Distribution Indicating the Existence of an Outlier Value. In These Situations, there may be a Case for Excluding the Outlier from Further Investigation.

When examining the connection between multiple datasets, the goal is to comprehend how these datasets combine and influence each other. This interrelation is referred to as correlation and may be either positive or negative, signifying whether the variables in question are helpful or counteractive towards each other. An effective way to visualize this is by employing a scatterplot. For data ranking, the most straightforward approach involves using a bar chart, which consists of a series of bars representing the progression of a variable. There are 4 main kinds of bar charts available: horizontal bar charts, vertical bar charts, group bar charts, and stacked bar charts.

Other Charts

Figure 11: A Varity of Charts are available for Conducting Data Pattern Analysis for Visual Purpose, as Shown in this Image.

Tables

When working with tabulated data, it's common to categorize or group the data into ranges. Sorting and filtering are widely used tools to facilitate the organization of data. Sorting involves arranging data in a particular order, while data filtering allows less relevant information to be concealed, enabling users to concentrate solely on the data of interest to them.

Potential Issues with Data Sets

Searching for data patterns becomes futile if the data itself is inappropriate. Therefore, it's crucial to evaluate the source and representativeness of the data, including its adequacy in terms of size. To ensure data suitability, it's essential to assess the following aspects:

1. Is the information accurate and reliable?
2. Are the data pertinent to the situation?
3. Does the data accurately represent the circumstances for which it is being used?
4. Is the information reliable throughout?
5. Was the definition used for all the data that was gathered?
6. Does any part of the data require adjustments to maintain consistency with historical patterns?
7. Does the data cover an appropriate time period?
8. Has a sufficient amount of data been included?

Machine Learning

In the realm of data pattern recognition, machine learning offers a more advanced approach. Machine learning algorithms have the capability to learn from data, and once optimized, they can autonomously identify patterns, even when they are only partially evident. While this process involves recognizing familiar patterns, the recognition occurs from various perspectives and angles, showcasing the valuable sophistication provided by machine learning (5).

VII. Summary

This article has explored some straightforward data representation tools, in addition to techniques for data capture and organization, as well as methods for examining data over time, assessing correlations, and understanding data distributions. It's important to note that there are many other techniques, and more intricate inquiries can be pursued. The aim here was not to provide a comprehensive guide but rather to offer a few examples for those embarking on their data review journey. In doing so, the emphasis has been on visually representing data rather than conducting an in-depth statistical analysis. Often, a visual representation can reveal significant insights about the data's shape and characteristics. This may suffice for the current inquiry, or it may serve as a preliminary step toward more extensive statistical assessments.

References

[1] F. Afrati, A. Gionis, and H. Mannila. Approximating a collection of frequent sets. In*Proc. ACM SIGKDD*, 2004

[2] Ajani, K., Lee, E., Xiong, C., Knaflic, C. N., Kemper, W., Franconeri, S. (2021). Declutter and focus: Empirically evaluating design guidelines for effective data communication. *IEEE Transactions on Visualization and Computer Graphics*. https://doi.org/10.1109/TVCG.2021.3068337

[3] Ancker, J. S., Senathirajah, Y., Kukafka, R., Starren, J. B. (2006). Design features of graphs in health risk communication: A systematic review.*Journal of the American Medical Informatics Association*, 13(6), 608–618

[4] Chance, B., delMas, R., Garfield, J. (2004). Reasoning about sampling distributions. In Ben-Zvi, D., Garfield, J. (Eds.),*The challenge of developing statistical literacy, reasoning, and thinking* (pp. 295–323). Springer.

[5] C. M. Velu and K. R. Kashwan, "Visual data mining techniques for classification of diabetic patients," 2013 *3rd IEEE International Advance Computing Conference* (IACC), 2013, pp. 1070-1075, doi: 10.1109/IAdCC.2013.6514375.

[6] Yang F, Harrison L T, Rensink R A, Franconeri S L, Chang R. Correlation judgment and visualization features: a comparative study. IEEE Transactions on Visualization and Computer Graphics, 2019,25(3): 1474–1488

[7] Giovannangeli L, Bourqui R, Giot R, Auber D. Toward automatic comparison of visualization techniques: application to graph visualization. Visual Informatics, 2020, 4(2): 86–98

[8] Liu Y, Zhang W, Wang J. Source-free domain adaptation for semantic segmentation. In: Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, 1215–1224