Data Visualization: Trends and Patterns

Ashish Gupta
Research Scholar, Dept. of CSE
Rabindranath Tagore University, Bhopal, India
guptaashishnitm@gmail.com
Dr. Sanjeev Kumar Gupta
Dean, Engineering
Rabindranath Tagore University, Bhopal, India
sanjeevgupta73@yahoo.com
Dr. Pritaj Yadav
Associate Professor, Dept of CSE
Rabindranath Tagore University, Bhopal, India
yadavpritaj@gmail.com
Dr. Deepak Gupta
Associate Professor, Dept of CSE
Institute of Technology and Management, Gwalior, India
deepak.gupta@itmgoi.in

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.

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 data (such as yield or critical process parameters), microbial counts, and categorization of deviations, among other factors. Data visualization serves as a valuable and universally understandable means for pattern analysis (1).

This article introduces several approaches for gaining insights from data using straightforward visual tools. While these tools are uncomplicated, they excel in conveying important points with clarity (2).

WHAT IS A DATA PATTERN?

When we attempt to break down a problem, one essential aspect is the search for discernible patterns within the generated data. Patterns are essentially similarities or shared characteristics that meet specific criteria. Complex pattern recognition is a fundamental concept in computer science and can also be applied to data sets using software tools like MS Excel or Minitab.

A data pattern within a dataset is essentially a sequence of data points that repeat in a recognizable manner. This recognition can be based on the historical data being analyzed or on data that exhibits similar characteristics. The simplest forms of 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 between two sets of numbers.

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 series models assume that the direction a chart takes is primarily related to its own historical patterns, while causal models assess the relationship between other influencing factors and the data under consideration.

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 observations gathered at a specific point in time, such as a series of tests conducted on a single in-process sample.

Time series data: These are collected over successive time intervals. For instance, it could involve a series of in-process samples examined 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-based data, there are typically four overarching pattern types: horizontal, trending, seasonal, and cyclical (3). Conversely, with cross-sectional data, the emphasis is more on extracting information and identifying patterns within individual events.

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 growth or decline within the time series over an extended duration. Line charts are particularly well-suited for visualizing continuous 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 measured over a period of time 1.

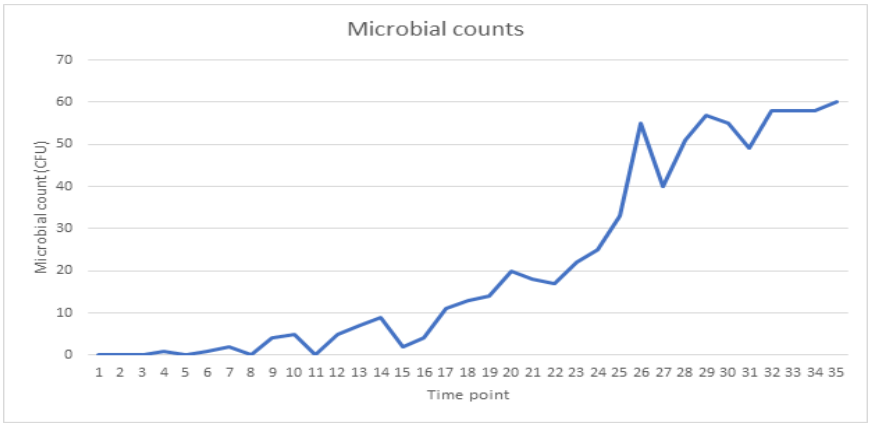

Fig 1: Data were analysed over time for microbial counts.

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

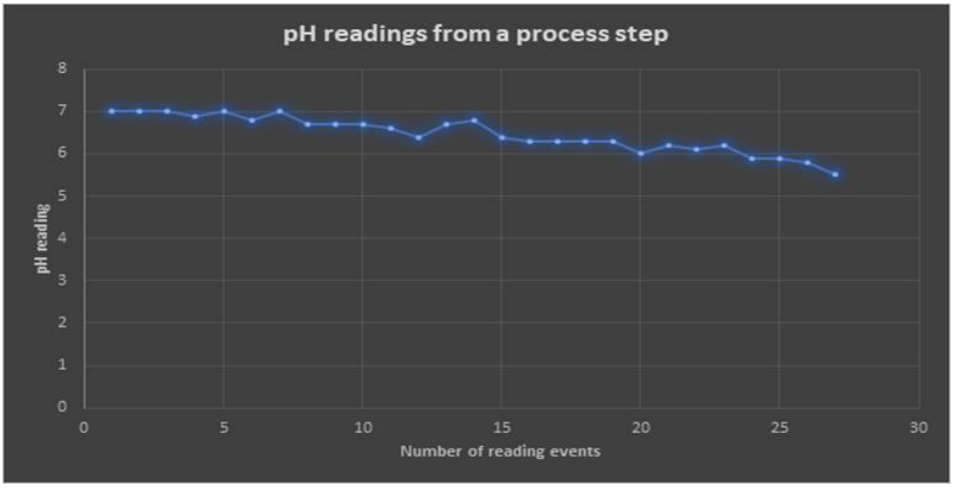

Fig. 2: Data analysis for pH readings over time

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

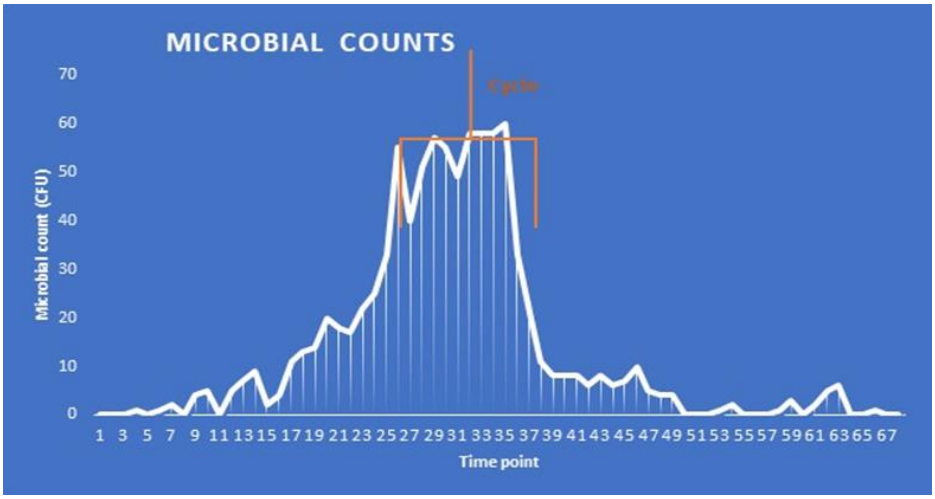

Fig. 3: A data cycle is focused on (in regard to microbial counts)

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

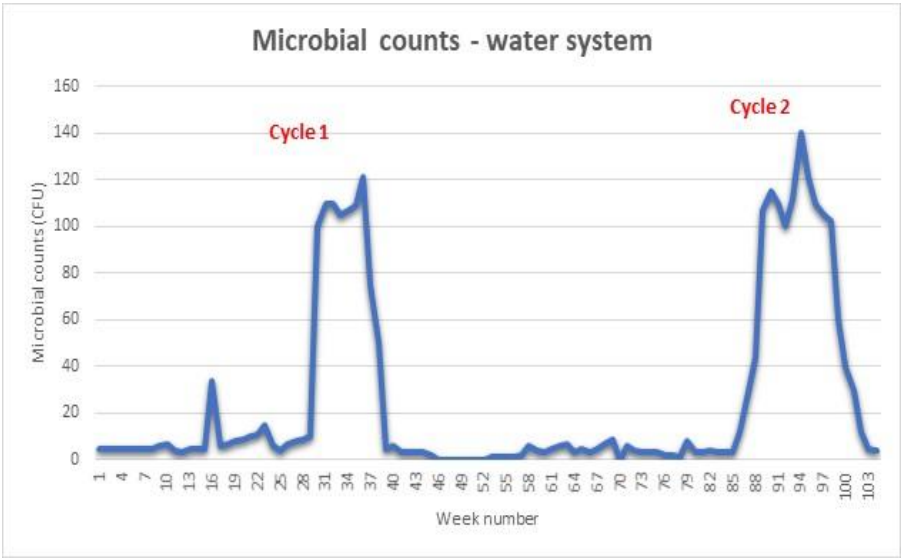

Fig 4: Example of a seasonal cycle for microbial counts that is repeated over time

In this case, two distinct cycles emerge over a span of two years, coinciding at roughly the same 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 maintenance reasons. Often, such cyclical patterns exhibit regularity, although their durations can vary.

These cyclical patterns can transition into the seasonal component, characterized by a repetitive pattern that recurs year after year. When data collected over time exhibit fluctuations around a constant level or mean, 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 an extended period, they would be classified as having a horizontal pattern.

EXPLAINING DATA PATTERNS

When evaluating collected data, it's valuable to provide descriptors that help characterize the data:
Are the data random, where successive values of a time series lack any discernible relationship?
Do the data exhibit a trend, indicating they are nonstationary?
Are the data stationary or horizontal?
Do the data display seasonality?
Using these descriptors, a series that fluctuates around a consistent level without showing growth or decline over time can be termed "stationary." Consequently, a stationary time series maintains constant basic statistical properties, such as mean and variance, as time progresses. Conversely, a series that includes a trend can be categorized as "non-stationary.**"**

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 throughout a study to make informed decisions or adjustments regarding design and study variables while maintaining experimental control and achieving improved outcomes. Additionally, it can aid in assessing data for normality or other characteristics before selecting the appropriate statistical analysis tool.

Patterns within data distribution are typically described in terms of four key aspects: center, spread, 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 graphic display where roughly half of the observations fall on either side. In the chart below,

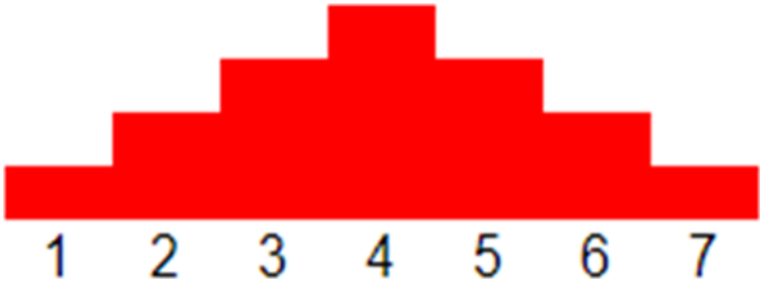

Fig. 5: Data demonstrating a centralised distribution

This is also coincidental with symmetry. When it is graphed, a symmetric distribution can be divided at the center so that each half is a mirror image of the other. Number of peaks. Distributions can have few or many peaks. Distributions with one clear peak are called unimodal, and distributions with two clear peaks are called bimodal. When a symmetric distribution has a single peak at the center, it is referred to as bell-shaped.

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

Fig. 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 Figure 8. Conversely, if observations cluster closely around a single value, the spread is narrower, as evident in Figure 7.

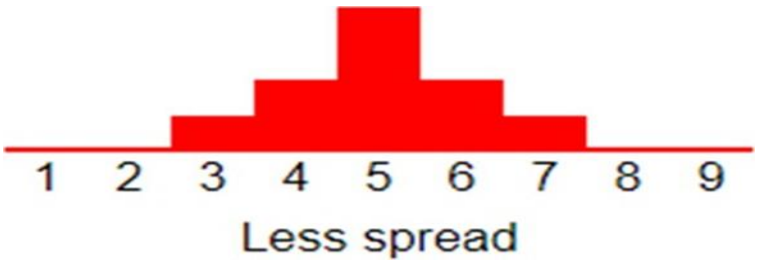

Fig. 7: Data that have a small spread

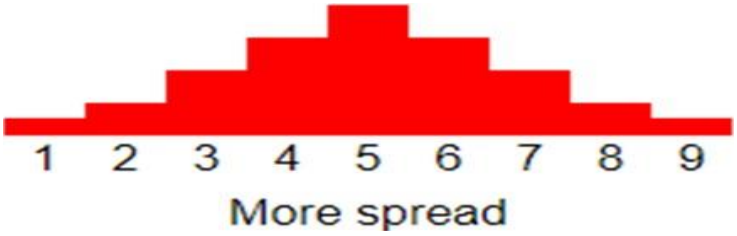

Fig. 8: Data with a comparatively wide spread

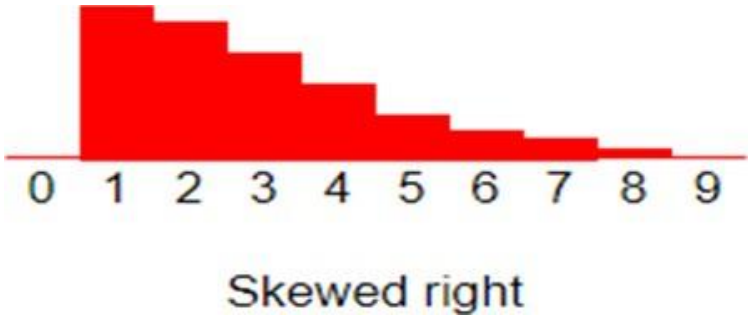

Fig. 9: Data that are skewed to the right, as if this were frequently the case for microbial data

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 other. Distributions with fewer observations on the right side (toward higher values) 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. These extreme values are referred to as outliers. As a general guideline, an extreme value is typically considered an outlier if it falls at least 1.5 interquartile ranges below the first quartile (Q1) or at least 1.5 interquartile ranges above the third quartile (Q3). An example of such an outlier is depicted in Figure 10.

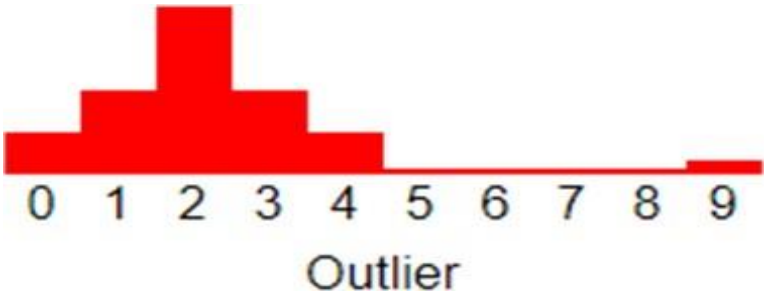

Fig. 10: Graph of the distribution indicating the existence of an outlier value. In these situations, there may be a case for excluding the outlier from further investigation.

OTHER CHARTS

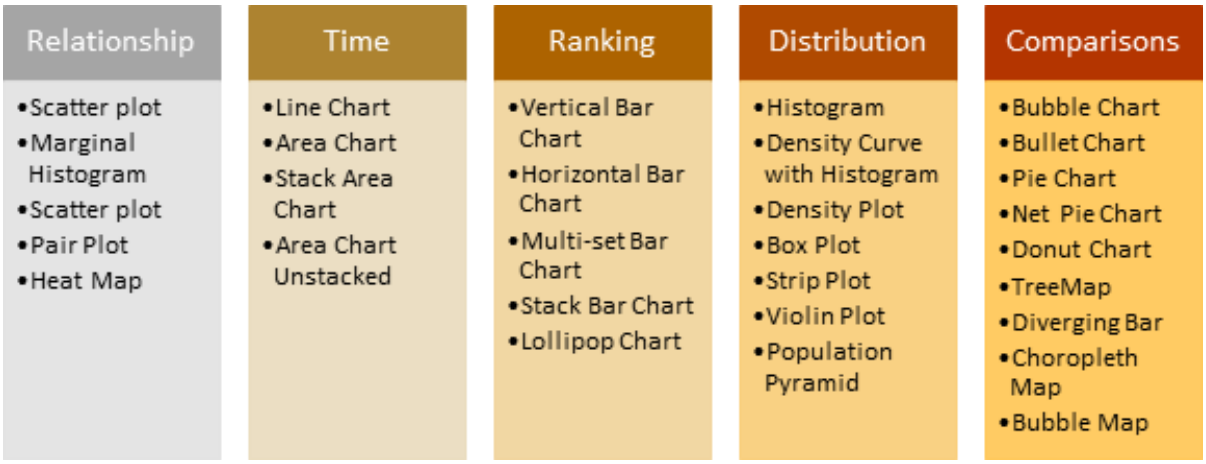

Fig. 11: A Varity of charts are available for conducting data pattern analysis for visual purposes, as shown in the image.

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 can be either positive or negative, signifying whether the variables in question are supportive 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 four main types of bar charts available: horizontal bar charts, vertical bar charts, group bar charts, and stacked bar charts.

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 unsuitable. 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:

Is the data reliable and accurate?

Is the data relevant to the context?

Does the data accurately represent the circumstances for which it is being used?

Is the data consistent throughout?

Was all of the data collected under the same definition?

Does any part of the data require adjustments to maintain consistency with historical patterns?

Does the data cover an appropriate time period?

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).

SUMMARY

This article has explored some straightforward data presentation tools, including 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 approaches, 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 riskcommunication: 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," 20133*rd 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