Proceedings of ICFTBA 2026 Symposium: Driving Organizational Strategic Change for Financial Gains
DOI: 10.54254/2754-1169/2026.BH38890

From Raw Data to Consumer Insights: A Review of Descriptive Statistics in Social Media Consumption Analysis

Haozhen Tang

School of Statistics and Data Science, Capital University of Economics and Business, Beijing, China
32024230025@cueb.edu.cn

Abstract. The vast amounts of user behaviour data accumulated on social media platforms provide valuable material for understanding consumer preferences, however, the raw data itself cannot be directly converted into business insights. Descriptive statistics transform unstructured data into comprehensible information through metrics such as the mean, median, frequency and standard deviation. Although this method is widely used in empirical research, a systematic review of its application in the context of social media consumption data is still lacking.This paper adopts a literature review approach to systematically examine five key aspects: the interpretation of core indicators, the application of three types of behavioural scenarios, the identification of high-potential individuals, the inherent limitations of the method, and standardised implementation processes. The study found that descriptive statistics are a fundamental tool for data exploration, capable of rapidly revealing data characteristics and analytical directions, thereby providing guidance for subsequent in-depth analysis, however, they cannot infer causal relationships or predict trends, and their limitations must be clearly defined in practice. Focusing on social media consumption data, this review systematically clarifies the interpretation of indicators, application boundaries and implementation processes for descriptive statistics. It fills a research gap in this field regarding indicator systems and scenario-based applications, and provides a methodological reference for the preliminary analysis of social media consumption data.

Keywords: Descriptive statistics, social media, consumer analysis, user behaviour, indicator system

1. Introduction

Consumers increasingly complete the whole journey—from product discovery to order placement—on social media, platforms such as Xiaohongshu, Douyin and Weibo have consequently accumulated vast amounts of user behaviour data [1]. These data span browsing history, engagement metrics and transaction information, providing an unprecedented opportunity to gain a comprehensive understanding of consumer preferences on a large scale [2]. However, raw data cannot automatically be transformed into actionable business insights, fields such as user identifiers, browsing duration and purchase amounts are merely discrete sets of numbers when not processed by a system [3].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2754-1169/2026.BH38890

Descriptive statistics serve as a fundamental analytical tool in this context, distilling complex datasets into interpretable summaries through metrics such as mean, median, frequency and standard deviation [4]. Virtually all empirical research in the fields of consumer behaviour and marketing begins with descriptive statistics to characterise the sample, however, the existing literature lacks a systematic review of their application specifically to the context of social media consumption data [5].

This paper employs a literature review approach to examine the subject from five perspectives: the definition of core indicators, typical application scenarios, methodological boundaries and limitations, and standardised implementation processes. It systematically clarifies the application value and scope of descriptive statistics in social media consumption analysis. It analyses the inherent limitations of the analytical methods and the biases inherent in social media data itself, whilst outlining descriptive statistics as a standardised process in the initial stage of business data analysis. This paper provides a methodological reference for researchers and practitioners, helping to standardise the use of descriptive statistics in social media consumption analysis and reduce common methodological misunderstandings.

2. Key indicators of descriptive statistics and their interpretation in social media consumption contexts

2.1. Measures of central tendency

The mean, also known as the arithmetic mean, is the most fundamental measure of central tendency [4]. In social media consumption analysis, the mean can be used to characterise metrics such as the average daily browsing time, average number of interactions and average order value. Its advantage lies in its computational simplicity and use of all available data, however, it is highly sensitive to outliers. A small number of high-spending users may skew the overall mean so that it no longer represents a 'typical user' [6]. For example, if a few consumers spend tens of thousands of yuan, the mean might reach several hundred yuan, whilst more than half of users actually spend less than 100 yuan. Therefore, using the mean on its own in a skewed distribution may be misleading.

$$\bar{x}=\frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

where, $\bar{x}$ is the mean, x_i is the i-th observation, and n is the total number of observations.

The median is the value situated in the middle when data are sorted, it is not influenced by outliers [4]. It is more suitable for describing the typical level of skewed data, such as typical user browsing time or spending amounts. In a right-skewed distribution, the median is usually significantly lower than the mean, and the difference between the two reflects the degree of skewness [5]. In practice, both the mean and the median should be reported.

The mode is the value with the highest frequency of occurrence and is applicable to categorical data [4]. It identifies the most common user attributes and behavioural patterns, such as the age group with the highest proportion, peak activity periods, and the most popular content types. Its advantage lies in its intuitiveness and ease of business interpretation, however, for continuous numerical data, it may not exist or be unique, and is typically used to complement the mean and median [5]. Figure 1 illustrates the typical positional relationship between the mode, median and mean in a right-skewed distribution, most social media consumption data exhibit this distribution pattern. The core indicators of descriptive statistics, their specific applications in social media consumption scenarios and their main limitations are summarised in Table 1.

DOI: 10.54254/2754-1169/2026.BH38890

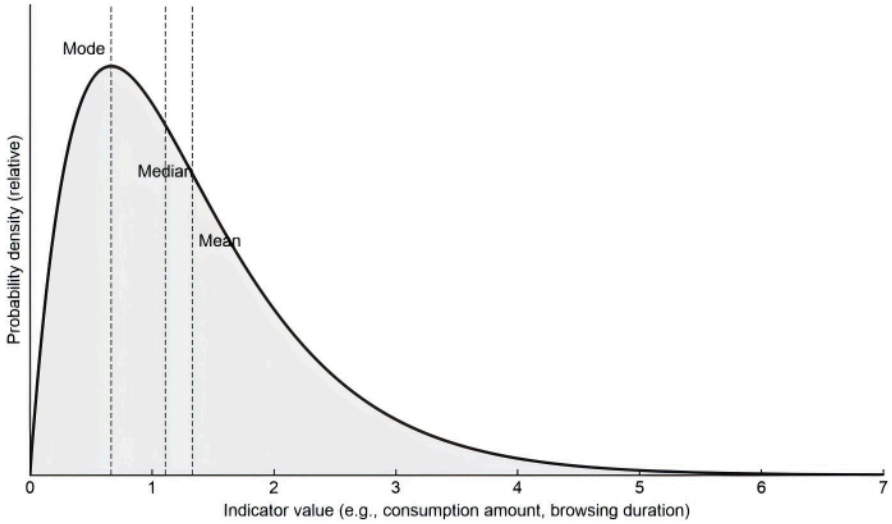


Figure 1. Positions of mode, median, and mean in a right-skewed distribution

Table 1. Summary of key descriptive statistics

Indicator category	Specific indicators	Examples of social media applications in consumer scenarios	Main limitations
Central Tendency	Mean	Average order value, average daily browsing time	Significantly affected by extreme values
	Median	Typical spend and typical browsing depth	Does not fully utilize all data
	Mode	Peak times, most popular content types	May not exist for continuous data
Distribution structure	Frequency/Percentage	User gender, age and geographical distribution	Describes only a single dimension
	Percentile	Definition of the top 20 per cent of high-value customers by spending	Depends on percentile cutoff choice
	Cumulative percentage	Pareto analysis identifies core consumer groups	Describes concentration only
Degree of dispersion	Range	Range of fluctuations in spending amounts	Relying solely on two extreme values
	Standard deviation	Assessment of the Stability of Interaction Volumes	Overestimated in skewed data

2.2. Distribution structure indicators

Frequency refers to the number of observations in a particular category, whilst percentage refers to the proportion of that category relative to the total number of observations [4]. Both are fundamental tools for describing the distribution of categorical data and are widely used to characterise users' demographic characteristics (gender, age, region, city tier, device type) and to describe the structure

DOI: 10.54254/2754-1169/2026.BH38890

of behavioural categories (content type preferences, mix of interaction types, distribution of consumption categories, among others) [2]. Their advantage lies in their intuitiveness and ease of understanding, however, they can only describe distributions along a single dimension and are unable to reveal correlations between dimensions.

A percentile is the value below which a specified percentage of observations in a data set falls [4]. It is a core tool for user segmentation and the identification of high-value user groups. For example, the P80 spending threshold indicates that 80 per cent of users fall below this value, those above the P80 constitute the top 20 per cent of high-value users in terms of spending, whilst the P90 browsing duration can be used to identify 'heavy users'. Percentiles overcome the 'averaging' limitations of the mean and median, enabling the precise identification of user tiers [5]. In practice, users are often categorised into four tiers—low, medium, high and very high—based on P50, P80 and P95, to facilitate differentiated operations.

The cumulative percentage is the running total of percentages after observations are sorted from largest to smallest, it is used to reveal the degree of concentration in a distribution [4]. Its most typical application is Pareto analysis (the 80/20 rule). Empirical research indicates that in the fast-moving consumer goods (FMCG) sector, the top 20 per cent of active buyers contribute on average approximately 73 per cent of brand revenue [7], in the online sales of digital brands, the top 20 per cent of customers contribute approximately 47 per cent of sales [8]. Figure 2 presents a typical Pareto analysis, where the cumulative spending curve clearly illustrates how consumption is concentrated amongst a small group of users. In social media e-commerce scenarios, the cumulative percentage curve can be used to precisely quantify the concentration of spending: if the top 10 per cent of users account for 65 per cent of spending, this indicates an extremely high level of concentration, and resources should be channelled towards the top performers, if the top 30 per cent of users account for only 60 per cent, the distribution is relatively dispersed. Pareto analysis provides a direct quantitative basis for the optimal allocation of resources in customer relationship management [3].

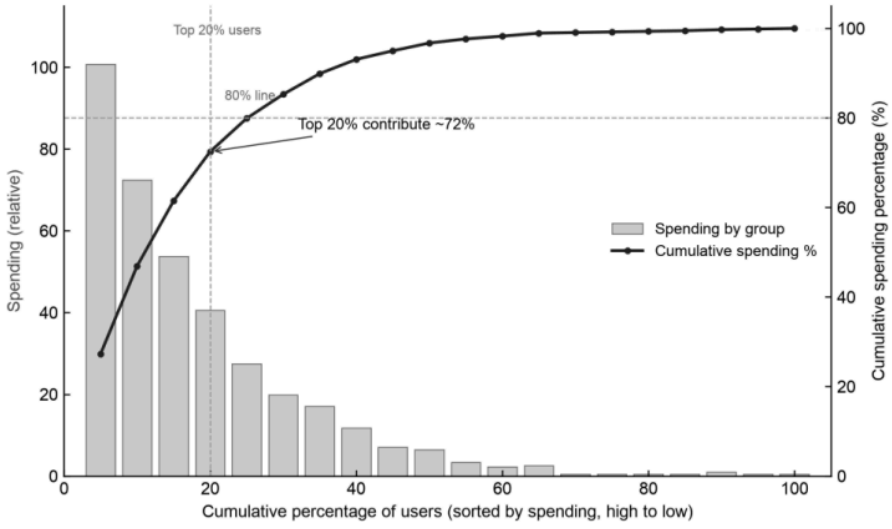


Figure 2. Pareto analysis of user spending concentration: cumulative spending percentage across users sorted by spending amount

DOI: 10.54254/2754-1169/2026.BH38890

2.3. Measures of dispersion

The range is the difference between the maximum and minimum values, and is the simplest measure of dispersion [4]. It provides a quick summary of variability in expenditure and browsing duration, however, as it relies solely on two extreme values and is highly susceptible to outliers, it is only suitable for preliminary exploration and should be used in conjunction with the standard deviation [5].

Variance and standard deviation measure the degree of dispersion of data points relative to the mean [4]. The standard deviation is more interpretable as it is consistent with the units of the raw data. In social media consumption analysis, the standard deviation can be used to assess the degree of dispersion in spending amounts, the volatility of engagement levels, and the stability of content performance, amongst other factors. A small standard deviation indicates consistent user behaviour and a highly representative mean, a large standard deviation indicates significant individual variation and a less representative mean [5]. For example, a low standard deviation in interaction levels for a series of content indicates stable performance and high predictability, conversely, an extremely high standard deviation suggests significant volatility, which may be due to occasional viral hits or inconsistent quality.

$$s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} \tag{2}$$

In social media operations, standard deviation can also be used to monitor content stability, assess consumption consistency and detect anomalies [1]. The application of standard deviation presupposes that the data follow an approximately symmetrical distribution, in the case of severely skewed social media consumption data, standard deviation may be overestimated due to outliers and should therefore be supplemented by dispersion measures based on percentiles, such as interquartile range, to support the analysis [4]. Measures of dispersion can only describe 'how large the differences are', they cannot explain 'why the differences exist'. Table 2 compares three typical application scenarios for descriptive statistics in social media consumption analysis, covering the core metrics for each scenario, the business questions they address, and their respective methodological limitations.

3. Typical applications of descriptive statistics in the analysis of social media consumption

3.1. Analysis of user browsing behaviour

Browsing behaviour reflects how users discover and consume content. The mean and median together describe the average daily browsing duration and typical browsing depth. As browsing duration typically follows a right-skewed distribution, the median often provides a more accurate reflection of the average user's behaviour, whilst the mean reflects the overall average, including heavy users [4]. Frequency analysis can identify preferences for content types and peak activity patterns. By counting the frequency with which users browse different content categories, it is possible to determine which types attract the most attention, analysing frequency distributions by the hour can reveal peak activity periods, helping to schedule content publication times to maximise reach [9]. For example, a cosmetics brand analysed the distribution of views across each hour of the week and discovered that 8.00 pm to 10.00 pm was a peak viewing period. Based on this, it scheduled the release of new product teaser videos for this time slot, resulting in an approximately 30 per cent increase in initial view counts. Descriptive statistics in browsing analysis answer the

DOI: 10.54254/2754-1169/2026.BH38890

questions 'what do users view' and 'when do they view it', but cannot explain why certain content attracts more views, the latter requires content analysis or experimental methods.

3.2. Evaluation of the effectiveness of user interaction behaviour

Interactions include 'likes', comments, shares and saves, and reflect the extent of user engagement with content. A combination of the mean and median is used to assess the 'day-to-day performance' of content: the mean shows the average interaction volume across all content, whilst the median filters out outliers caused by viral hits, reflecting the performance of typical content [4]. The mean and standard deviation are used in conjunction to assess the stability of content performance. A high mean and low standard deviation indicate consistently strong performance, a high mean with a high standard deviation suggests that results are driven by a few viral hits and that overall quality is unstable [5]. For example, after an educational account published 20 short videos in a row, it used the mean to compare average engagement levels across different topic areas and the standard deviation to assess the volatility of each area. Based on this analysis, the account eliminated topic areas with low average engagement and high volatility, focusing instead on topics with stable, high-average performance. Engagement rate (total engagement divided by reach or impressions) is another derived descriptive metric, widely used to compare content performance across accounts of different sizes. The formula for calculating the engagement rate is:

$$\text{Engagement rate} = (\text{Likes} + \text{Comments} + \text{Shares} + \text{Favorites}) \,/\, \text{Reach} \times 100\% \tag{3}$$

3.3. Consumer behaviour and the identification of high-potential individuals

Actual purchasing behaviour is the ultimate conversion metric. Key descriptive metrics include average order value, purchase frequency and repeat purchase rate. Whilst the mean is used for average order value, the median should also be reported to account for skewness, purchase frequency is commonly summarised using the mean and mode, and the repeat purchase rate is a percentage calculated based on frequency [4]. The identification of high-potential customers follows a three-step process: first, a baseline is established by calculating the mean, median and percentile benchmarks for spending and purchase frequency, second, a stratified analysis is conducted according to key percentile points (such as the top 20 per cent by spending), and finally, the demographic and behavioural characteristics common to high-tier users are analysed using frequency and percentage data [10]. A typical application is precision marketing: a certain online clothing retailer segmented users into four tiers based on annual spending and found that users in the P95 tier and above were predominantly women aged 25 to 35 who favoured sportswear and trainers. Consequently, the retailer sent targeted VIP previews of new products to this group, resulting in a repeat purchase rate approximately 20 per cent higher than that of standard promotions. Although this approach identifies customers with the highest current expenditure, it does not explain the factors driving their spending. It is therefore primarily descriptive rather than predictive and cannot, on its own, reliably identify which customers will become high-value customers in the future [5]. Causal attribution and prediction require inferential statistics or machine learning methods. A comparison of the core metrics, business questions and methodological limitations for the three typical application scenarios is shown in Table 2.

DOI: 10.54254/2754-1169/2026.BH38890

Table 2. Comparison of the three types of application scenarios

Application scenarios	Key indicators	Business-related questions answered	Limitations of the method
Browsing behaviour	Mean / Median, Frequency	What users watch and when they watch it	Unable to explain the reasons for the content's appeal
Interactive behaviour	Mean + standard deviation, frequency, engagement rate	Is the content consistent, and what is the audience's sentiment?	It is not possible to measure differences in the quality of interactions
Consumer behaviour	Average order value, purchase frequency, repeat purchase rate, percentile	Who are high-value customers?	Unable to explain consumer motivations, lacks predictive power

4. Limitations of descriptive statistics and bias in social media data

4.1. Inherent limitations of the method

The limitations of descriptive statistics are primarily evident in three areas. The most fundamental limitation is that it merely summarises and describes existing data, it cannot infer causal relationships. For example, whilst it may be observed that users who spend more time browsing spend more, this does not prove that longer browsing time leads to higher spending, there may also be reverse causality or the influence of a third variable [4]. Furthermore, it cannot predict future trends and can only summarise past and current conditions. Thirdly, descriptive statistics are sensitive to the choice of metrics, the same dataset may yield different conclusions depending on whether the mean or the median is reported, whilst omitting the standard deviation may obscure data volatility [5]. Furthermore, as descriptive statistics merely summarise existing data, they cannot identify whether systematic omissions or measurement biases exist in the data collection process, this constitutes a significant limitation to their application. Researchers should select metrics based on the characteristics of the distribution and report multiple measures when the data are skewed.

4.2. Bias in social media data

Social media data are subject to biases that cannot be detected or corrected by descriptive statistics alone [11]. Firstly, there is a representativeness bias: social media users are not equivalent to the general consumer population, there are structural differences in terms of age, gender, region and urbanisation level, with younger and urban users being over-represented, whilst older and rural consumers are under-represented. Secondly, there is a participation bias: users who actively share their consumption behaviour on social media may differ in their desire for self-expression or in their consumption patterns [11]. The 90-9-1 rule (i.e. approximately 90 per cent of users merely browse without participating, 9 per cent interact occasionally, and 1 per cent create the majority of content) indicates that the vast majority of user-generated content originates from a very small number of highly active users, the observed behaviour may therefore reflect this active minority rather than the broader user base. Thirdly, there is platform-specific bias: data from a particular platform (such as

DOI: 10.54254/2754-1169/2026.BH38890

Douyin) may not be generalisable to other platforms (such as Xiaohongshu) due to differences in user profiles and content culture. Descriptive statistics accurately summarise the given data, but cannot determine whether those data are representative. Researchers must take these biases into account when interpreting results to avoid over-generalising social media data to the entire consumer population [11].

5. Standardised procedures for descriptive statistics as a preliminary step in business analytics

As the starting point for business data analysis, descriptive statistics typically follow a standardised process. Step one: data cleaning and validation—removing duplicate values, handling missing values, and checking for outliers that may skew the results. Step two: variable classification—distinguishing between numerical variables (spending, browsing duration) and categorical variables (gender, content type) to select appropriate metrics. Step three: description of central tendency: for numerical variables, report the mean and median, for categorical variables, report the mode. Step four: distribution and dispersion analysis—use frequencies and percentages for categorical variables, use percentiles for user segmentation, and use standard deviation to measure dispersion (or the interquartile range for skewed data). Step five: cross-tabulation and preliminary comparisons—examine the relationship between two variables through grouped descriptive statistics (e.g. expenditure across different age groups). Step six: interpretation of results and setting the direction: summarise key findings, identify anomalous or interesting patterns, and determine which issues warrant further inferential analysis [3]. This process ensures that subsequent advanced analyses, including regression, clustering and predictive modelling, are built upon a solid understanding of the data's fundamental characteristics. Omitting descriptive statistics often leads to misinterpretation of advanced results, as researchers may fail to identify skewness, outliers or biased distributions that affect model performance [12].

6. Conclusion

This paper systematically reviews the application of descriptive statistics in the analysis of social media consumption. The review examines key metrics such as the mean, median, mode, frequency, percentiles and standard deviation, and explains the interpretation of each metric within the context of social media, it then explores the application of this methodology in three scenarios: browsing behaviour, the evaluation of engagement effectiveness, and the identification of consumer behaviour and high-potential customers, whilst also assessing the inherent limitations of the method—namely, that descriptive statistics cannot establish causal relationships, cannot make predictions, and are sensitive to the choice of metrics, it further discusses the sampling bias, participation bias and platform-specific bias embedded within social media data, finally, it proposes a six-step standardised process for utilising descriptive statistics as the initial stage of business data analysis.

The findings of this study indicate that descriptive statistics can concisely distil complex social media consumption data into an interpretable summary and provide necessary guidance for subsequent in-depth analysis, however, their limitations must be recognised. They answer 'what the data look like', rather than 'why these patterns exist' or 'what will happen next'. This review has several limitations. Firstly, the paper relies primarily on existing literature rather than conducting empirical analysis using actual social media datasets, the practical applicability of some of the recommended metrics requires further validation. Secondly, the scope of the study is limited to descriptive statistics itself and does not address how this method interfaces with specific inferential or machine learning methods in practice. Future research could conduct empirical case studies using

DOI: 10.54254/2754-1169/2026.BH38890

real social media consumption data, compare metric selection across different product categories and platform types, and explore how findings from descriptive statistics can be incorporated into specific predictive models to improve consumer insights.

References

[1] Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons, 53*(1), 59-68.
[2] Liu, Q., Wang, R., Pareti, M., Castellini, A., Viaggi, D., & Canavari, M. (2026). Big data in consumer behavior research: A systematic review of data sources, analytical methods, and research questions. *Journal of Marketing Analytics, 1*-23.
[3] Davenport, T. H., & Harris, J. G. (2007). *Competing on analytics: The new science of winning*. Harvard Business Review Press.
[4] Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J. D., & Cochran, J. J. (2020). *Statistics for business and economics* (14th ed.). Cengage Learning.
[5] Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Prentice Hall.
[6] Chen, P., & Sui, J. (2013). Statistical analysis of the characteristic behaviors of Weibo users based on individual attributes. *Knowledge Management Forum*, (3).
[7] Kim, B. J., Singh, V., & Winer, R. S. (2014). The Pareto rule for frequently purchased packaged goods: An empirical generalization. *Marketing Letters, 25*(4), 393-404.
[8] Dawes, J. G., & Nenycz-Thiel, M. (2025). Sales concentrations of digital brands. *Journal of Marketing Management, 41*(1-2), 1-24.
[9] Pletikosa Cvijikj, S., & Michahelles, F. (2013). Engagement in social media platforms: A case study of branded pages. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 554-558).
[10] Wei, M., Zhang, H., Liu, Y., & Xu, H. (2019). Research on high-impact user profile of social media based on multi-dimensional attribute fusion. *Documentation, Information & Knowledge*, (5), 73-79.
[11] Olteanu, A., Castillo, C., Diaz, F., & Kici, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data, 2*, 13.
[12] Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165-1188.