Introducing num_dist: A Revolutionary Visualization R function for Enhancing Data Science Explorations

Yuqing Xue

School of Public Health and Health Sciences, Tianjin University of Traditional Chinese Medicine, Tianjin, China

Email: drxueyuqing@163.com

Abstract. Exploratory Data Analysis (EDA) is a fundamental practice in data science, enabling practitioners to uncover patterns, identify anomalies, and generate hypotheses without preconceived assumptions. However, traditional EDA tools often fall short in handling datasets with skewed or heavy-tailed distributions, where key insights are often hidden in the tail regions. *num_dist*, a function within the developing SmartVisual R package, addresses this gap by providing an interactive, Shiny-based tool that focuses specifically on the tail behavior of numerical distributions. By allowing users to zoom into the tail regions, adjust data ranges dynamically, and compare subgroups, *num_dist* enhances the exploration of outliers, rare events, and extreme values. Through real-time statistical summaries and intuitive faceting, it empowers data scientists to generate deeper insights, refine hypotheses, and make data-driven decisions. This paper demonstrates how *num_dist* promotes a more effective and efficient approach to EDA by facilitating a detailed examination of critical data segments often overlooked by traditional methods. With its focus on interactive exploration and hypothesis generation, *num_dist* serves as a valuable tool in modern data science, particularly for datasets with complex distributions and extreme values.

Keywords: Exploratory Data Analysis, Data Visualization, heavy-tailed, RShiny

1 Introduction

Exploratory Data Analysis (EDA) is widely recognized as the foundational step in the data science workflow [1]. It serves as the first line of interaction with data, allowing practitioners to uncover hidden patterns, detect anomalies, and develop hypotheses without the constraints of predefined assumptions. EDA offers critical insight into the nature of the data, helping data scientists identify potential variables of interest, relationships, and even biases in their models. However, despite its power, traditional EDA tools often fall short in handling datasets with complex distributions, especially those that contain skewed numerical variables or extreme values in the tails of distributions.

In many real-world datasets, important insights are often hidden in the tail regions of numerical distributions. For example, in finance, the analysis of tail events such as

N. B. Abdul Ghani et al. (eds.), *Proceedings of the 2025 6th International Conference on Big Data and Social Sciences (ICBDSS 2025)*, Advances in Computer Science Research 126,
https://doi.org/10.2991/978-94-6239-598-5_19

stock market crashes or extreme price movements is critical for risk management. Similarly, in healthcare, rare but critical events such as the onset of a disease in an unusual population group often reside in the tail of the distribution. Unfortunately, failing to put enough attention on this critical issue, traditional visualization tools like histograms or box plots, provide a broad overview of the data but do not allow for in-depth exploration of these rare or extreme events.

This is where *num_dist*—a function within the developing SmartVisual R package—comes into play. While *num_dist* focuses on the tail regions of distributions, incorporating advanced statistical methods such as Extreme Value Theory (EVT) or the Hill estimator could further enhance its utility for rigorous statistical inference. This theoretical grounding would allow for quantitative tail behavior analysis, distinguishing extreme values from noise with more precision.

By focusing specifically on the tail regions of numerical distributions, *num_dist* offers a more effective method for uncovering these important insights. The function is designed with interactivity and flexibility in mind, allowing data scientists to zoom in on specific parts of a distribution, analyze outliers, and generate hypotheses based on real-time exploration of data.

In this article, we explore how *num_dist* enhances the process of data exploration, promotes deeper insights, and improves objective hypothesis generation, particularly when dealing with skewed distributions, outliers, and rare events that specifically may carry critical information of understanding the data.

2 Key Features of *num_dist*

RShiny is a powerful framework for building interactive web applications directly within R, enabling users to create dynamic and user-friendly data visualizations. It allows seamless integration with various R packages, including plotly, to create rich, interactive visualizations that enhance user engagement and data exploration. RShiny supports the development of complex, real-time dashboards and web apps, making it a valuable tool for data scientists and analysts seeking to present their results in an accessible and interactive format [3,4] .

To deliever truly powerful tools, we use RShiny to build num_dist. It stands out from conventional EDA tools due to its specialized focus on tail analysis. It leverages interactivity and dynamic features to provide users with a comprehensive understanding of the numerical variables in their dataset. Below are the key features of num_dist that make it a powerful tool for data scientists.

2.1 Focus on Tail Behavior

In many datasets, especially those in domains such as finance, health, or environmental science, the most significant events often occur in the tail ends of the distribution [2]. Traditional histograms or summary statistics, such as the mean and standard deviation, offer little insight into these critical areas: signal and noise can switch under the scenario of heavy-tailed distribution. Means and standard deviations can be totally useless.

Fully aware of that, *num_dist* is designed to specifically address this limitation by providing a detailed focus on the tails of the distribution. Users can interactively zoom in on the tail regions to explore the behavior of extreme values in more depth.

This focus on tail behavior allows users to detect anomalies that might otherwise be overlooked in the bulk of the data. Whether it's spotting a financial crash, identifying rare diseases, or zooming in extreme flights delay, *num_dist* offers a unique advantage for data scientists looking to understand the extreme ends of the distribution.

2.2 Dynamic Range Selection

Utilizing RShiny's powerful interactive features, one of the most powerful features of *num_dist* is its dynamic range slider, which allows users to interactively adjust the range of data to be analyzed. This slider enables data scientists to zoom in on specific parts of the distribution in real-time, focusing on any range of values that might be of particular interest. For instance, in a dataset of flight delays, a data scientist might want to zoom into the extreme right tail of the distribution to examine the most severe delays, without manually filtering out the rest of the data.

By adjusting the range, users can explore different segments of the distribution and uncover patterns or outliers that are specific to particular ranges of values. This flexibility reduces the need for pre-defined assumptions and helps ensure that insights emerge directly from the data itself, rather than from arbitrary thresholds set in advance. This is a schematic strategy that is essential to generate iterative questions that effectively leading to deeper questions of better understand the data.

2.3 Grouping and Faceting

num_dist also allows users to group data by categorical variables (such as airline, city, or product category) and facet the data according to these groups. This feature is particularly useful for understanding how the tails of numerical distributions vary across different subgroups. For example, in the analysis of flight delays, *num_dist* can be used to compare the tail behaviors of delays across different airlines, airports, or times of day. This helps data scientists identify subgroups that might exhibit unusual or extreme patterns.

In the context of today's Data Science projects, high granularity / high resolution exploratory data analysis is gaining much more relevance, one might say it is necessary to generate complex hypotheses of the dataset. Faceting enables a more granular analysis of the data, providing insights into how specific factors, such as the type of product or the operational environment, influence the tail behavior of a distribution. This ability to compare subgroups in real-time enhances the depth of the analysis and helps identify factors that contribute to extreme outcomes with great ease and accessibility for much more wider audiences and practitioners, where in many scenarios, the extreme outcome identification is not systematically addressed by current EDA tools.

2.4 Statistical Summaries

Accompanying the interactive visualizations, *num_dist* provides real-time statistical summaries that update dynamically as the user interacts with the data. These summaries include key metrics such as the mean, median, standard deviation, minimum, maximum, and sample size. For example, when zooming in on a particular range of data, the statistical summaries provide updated values for the mean and median, helping users interpret how extreme values or outliers affect the overall distribution.

The inclusion of dynamic statistical summaries enhances the interpretability of the visualizations and allows users to make more informed decisions about the significance of the observed patterns. These summaries are crucial for confirming or disproving hypotheses and for providing a more robust understanding of the data.

3 How *num_dist* Promotes Data Science

num_dist is not just a visualization tool—it's a transformative resource that enhances the entire data science workflow. Below are some key ways that *num_dist* supports data scientists in their efforts to uncover deeper insights, detect rare events, and generate hypotheses without bias or relying on pre-assumptions.

3.1 Enhancing Tail Analysis

The core benefit of *num_dist* is its ability to facilitate focused analysis of the tail regions of distributions. Many critical insights in data science lie in the extreme ends of distributions, and *num_dist*'s functionality makes it easier to identify and understand these tail events. For example, in finance, the occurrence of extreme market events, such as flash crashes or sharp market rallies, can have a profound impact on risk management. In healthcare, detecting rare diseases or unusual patient outcomes can lead to breakthroughs in treatment or prevention strategies.

num_dist's interactive zooming capabilities allow data scientists to focus on the rare but important extremes of the data, making it easier to spot outliers, uncover hidden patterns, and gain insights that would be totally overlooked or / and difficult to detect with traditional static visualizations.

3.2 Uncovering Outliers and Rare Events

Outliers are often seen as problematic data points that should be excluded from analysis. However, in many cases, these outliers carry important information about the underlying processes or phenomena. For example, in an analysis of customer purchasing behavior, an outlier might represent a rare but significant purchase event that is critical for understanding a niche market. *num_dist* helps users identify and explore these outliers by focusing specifically on the tail regions of the distribution.

The ability to zoom into the tails and examine outliers in detail enables data scientists to evaluate whether these extreme values represent genuine signals or whether they

should be treated as noise. The decision of directly categorizing the "far-away" data points as outliers and deleting them without even nearly enough consideration can be catastrophic when the so called "outlier" may just be the crucial signal tha shape the intrinsic feature of the model. This insight is critical for decision-making, especially when outliers significantly influence model results.

3.3 Facilitating Hypothesis Generation

Asking correct questions is the key. In data science, hypothesis generation is the key step leading to deeper and more precise questions, which is the powerful and necessary iterative process that requires a deep understanding of the data. *num_dist* facilitates this process by providing interactive tools that allow users to dynamically explore different ranges of data and refine their hypotheses in real-time. For example, a data scientist analyzing flight delays may hypothesize that delays vary by airport. With *num_dist*, they can zoom into the tail of the delay distribution and compare delays across airports to refine their hypothesis, identify and correct pre-assumptious bias, and identify trends or anomalies.

This flexibility promotes an open-ended exploration of the data, where hypotheses can evolve as new insights are uncovered. *num_dist* enables data scientists to test multiple hypotheses without the need for repetitive coding or manual data manipulation, which often poses as the very hurdle to prevent analyst doing proper and thorough exploratory data analysis.

3.4 Providing Intuitive Group Comparisons

By allowing users to group data by categorical variables and facet by these groups, *num_dist* simplifies the process of comparing the tail behavior of numerical variables across different subgroups. This is especially important in scenarios where the influence of categorical variables on numerical outcomes is not immediately obvious. For example, in analyzing delays across different carriers, *num_dist* allows users to compare the tail behaviors of delays for major airlines versus smaller carriers, helping to uncover patterns that would otherwise remain hidden.

These group comparisons make it easier to understand how different factors influence the distribution and extreme behavior of numerical variables, leading to more actionable insights.

3.5 Reducing the Need for Repetitive Coding

Traditional EDA tools often require a significant amount of repetitive coding to filter and manipulate data. For example, to focus on specific ranges of a distribution, data scientists may need to write custom code to filter data subsets and generate new visualizations. *num_dist* eliminates this need by providing an interactive interface that automatically updates as users adjust the range slider or select different options. This feature greatly reduces the amount of code needed for analysis, allowing data scientists to focus more on interpretation and less on coding.

4 Evaluation of num_dist

Compared to point and click software, advantages of using R for EDA is its transparency and reproducibility [5]. Current R-based EDA or visualization packages or approaches [6-10] exhibit clear limitations, especially when handling complex data types. These tools perform well with simple datasets, like uniformly distributed categorical or symmetric numerical data, providing basic analysis. However, they struggle with more intricate data challenges such as skewed numerical distributions, rare categories, and multi-variable associations. For example, these packages often lack the ability to focus on tail-end behaviors in skewed data or to quantify the impact of rare categories, often leading to missed insights. Additionally, their static nature and limited interactivity make it difficult to explore complex relationships in real-time or across different groupings.

num_dist elevates exploratory data analysis by moving beyond the constraints of conventional tools. It is engineered for the demands of contemporary data work, providing a more dynamic and insightful way to understand numeric variables.

- Dynamic, On-the-Fly Investigation: Instead of writing new code for each query, users can explore data instantly with interactive controls like sliders, filters, and split-view options. This immediacy was key in identifying a bimodal speed pattern in longer flights simply by adjusting a slider to examine the data's extremes, dramatically speeding up the process of forming and checking hypotheses.

- Spotting the Unusual and Infrequent: The toolkit includes specialized functions to bring outliers and sparse categories into sharp relief. Using intelligent scaling (like square-root or log axes), it ensures that rare but significant events—such as the notable delays of a specific airline—are visible. This is crucial for analyzing real-world data where the most important insights often lie in the tails of a distribution.

- Designed for Effortless Use: Built with the user in mind, the package minimizes setup time through automatic variable recognition and intuitive controls. Its resilience is built-in, with safeguards that manage common data issues like outliers and missing values, ensuring dependable performance across various analytical tasks.

- Adaptable and Extensible Framework: The system is versatile, accommodating different data types and analytical questions. Its modular architecture means it can be easily expanded for new analyses, such as time-series, and can dissect data through multiple groupings for thorough investigation.

- Polished, Cohesive Visual Output: Every chart is rendered with a distinctive, professional aesthetic—featuring a dark theme with high-contrast elements. This visual uniformity not only creates publication-ready graphics but also makes complex patterns, like multi-peak distributions, easier to perceive and present to others.

These strengths collectively enable *num_dist* to address the limitations of traditional EDA tools, providing a seamless and powerful exploration experience.

We compare the effectiveness of *num_dist* with other approaches used for tail analysis:

Log Transformations

Logarithmic transformations are often used to address right-skewed distributions, where most of the data are clustered at the lower values, and extreme values lie in the

tail. Applying a log transformation can reduce the influence of extreme values and make the data more symmetric.

Comparison: *num_dist* provides an interactive tool that allows users to zoom into specific tail regions of the distribution. In contrast, log transformation is a static preprocessing technique applied to the entire dataset. While log transformation makes the tail behavior more manageable by compressing extreme values, it does not offer the flexibility of focusing on specific tail regions dynamically like num_dist does. Therefore, *num_dist* enables more nuanced insights into the extremes of the data without altering the original data distribution.

Quantile-Quantile (Q-Q) Plots

Q-Q plots are a standard method for comparing the quantiles of a dataset with those of a reference distribution (often a normal distribution). This method is useful for detecting deviations from the assumed distribution, especially in the tails.

Comparison: While Q-Q plots can visually indicate whether the data's tail deviates from a theoretical distribution, they do not allow for dynamic exploration of specific tail regions. *num_dist,* on the other hand, provides an interactive interface to focus on different parts of the data's tail, making it easier for data scientists to detect anomalies and extreme values that may not be immediately obvious in a Q-Q plot.

Outlier Detection Methods

Existing outlier detection methods, such as the outliers or extremevalues R packages, are commonly used to identify data points that fall far from the mean. These methods typically rely on predefined thresholds, statistical assumptions, or parametric models to classify outliers.

Comparison: *num_dist* offers a more flexible and user-driven approach, allowing the exploration of the tails of a distribution without relying on predefined thresholds. Unlike traditional outlier detection methods, which may fail to capture nuanced extremes in skewed distributions, *num_dist* dynamically highlights extreme values in their original context, making it more intuitive and responsive to the dataset at hand.

Extreme Value Theory (EVT) and Peak-Over-Threshold (POT) Models

Extreme Value Theory (EVT) focuses specifically on modeling the behavior of extreme values, often using tools like the Peak-Over-Threshold (POT) model, which identifies values exceeding a certain threshold and models their distribution.

Comparison: While *num_dist* provides an interactive, user-driven tool for exploring tail behavior, EVT and POT models require specific statistical methods and assumptions about the tail's behavior. *num_dist* could be further enhanced by incorporating EVT methods, which would allow it to not only visualize but also statistically model extreme values. A comparison with EVT methods would help illustrate the benefits of *num_dist* in terms of both exploration and the potential for more formal statistical analysis.

5 Practical Application: Case Study with nycflights13 Dataset

Let's consider a practical application of *num_dist* using the *nycflights13::flights* dataset, which is a popular dataset presented in R for Data Science as the exmplified dataset for data transformation, contains detailed information about over 336,000 flights, including variables like departure and arrival delays, distances, and other flight attributes. The dataset has several skewed distributions, particularly with variables like *arr_delay* (arrival delays), which exhibit heavy tails.

5.1 Extreme Delays

Using *num_dist*, we can zoom into the right tail of the *arr_delay* distribution to focus on extreme delays, the bin heights of which are always too small to be properly presented on the typical histogram or density plots. This zoomed-in view helps identify flights that experienced unusually long delays—valuable for understanding rare, yet impactful events. By zooming in on extreme values, we might uncover systemic issues, such as weather-related disruptions, that are influencing a large number of flights with the function of group variable selection for study the associations between categorical variable and numeric variables. These critical feature is not in the scope of this manuscript.

5.2 Flight Distance and Delays

Faceting by flight distance allows us to analyze how delays behave for short, medium, and long-haul flights. *num_dist* might reveal that longer flights tend to have higher delays, possibly due to the greater number of factors (weather, air traffic, etc.) that affect longer journeys.

Also, combining faceting with stacked boxplots and 3D plots yield a powerful and effective comprehensive visualization strategy to intuitively look into potential interactions between two categoricals when the responsible variable is numeric.

5.3 Outlier Detection

Outliers in flight delay data can reveal significant operational inefficiencies, such as certain flights or airports being consistently delayed. *num_dist'*s interactive zooming helps data scientists identify these outliers and investigate their root causes, such as poor scheduling or external operational disruptions.

6 Discussion

num_dist represents a significant advancement in the practice of Exploratory Data Analysis by focusing specifically on tail behaviors and rare events, with the speical focus on semi-automatically producing assumption-free hypothesis as practitioners explore through dataset using the function. Its interactive nature, coupled with dynamic range

selection and group comparisons, empowers data scientists to explore complex datasets more efficiently and effectively. By facilitating the exploration of outliers and rare events, *num_dist* helps uncover valuable insights that traditional tools might miss.

It is crucial to realize that Data Science projects today requires a different set of EDA tools to gain data literacy than what normally statsitsicans familiar with, and *num_dist* was built highlighting this requirement:

- Interactivity Feature: The ability to dynamically adjust ranges and filters was critical for uncovering pattern, emphasizing the need for real-time exploration tools in EDA. This suggests that future tools should prioritize user-driven interactivity.

To validate the efficacy of num_dist in real-world data science workflows, we propose conducting a usability study involving data practitioners. Feedback regarding interface intuitiveness, workflow integration, and time-to-insight would provide concrete evidence for the tool's effectiveness and improvements in EDA efficiency. Gathering such feedback will be an important task conducted in our further research.

- Accesible and Simplicity: *num_dist* garuantees intuitive interface and automatic and dynamic functions enables ease of use, a key consideration from users' perspective that are often overlooked by designers of EDA tool with similar functions in general. For instance: 1. *num_dist*'s automatic bin-width selection mechanism ensures optimal representation of the distribution of the interested numeric variable, hence maximizes user's intuitive ability to gain direct impression from the visualization; 2. *num_dist* provides immediete responsive descriptive analysis for each interactivity - result tables positioned directly under the visualization that greatly reduces the common headache of redundancy in the analysis process in result reporting, making it practically useful to a much wider audience. Futhermore, the idea of looking at both the result and the corresponding summary at the same time provides a unique oppurtunity for further enhanced hypothesis inspiring and problem iteration.

For seamless adoption, num_dist is designed to integrate effortlessly with other widely used R packages like ggplot2, dplyr, and shinyWidgets. Clear documentation on its API, extensibility, and customization options will ensure that num_dist fits into existing data analysis pipelines and empowers advanced users to extend its functionality.

7 Conclusion

The *num_dist* function represents a significant advancement in Exploratory Data Analysis (EDA), particularly for datasets exhibiting complex, skewed, or heavy-tailed distributions. Traditional EDA tools often overlook the critical insights present in the tail regions of distributions, especially in real-world datasets where rare events or outliers can provide valuable information. By focusing specifically on these areas, *num_dist* empowers data scientists to explore and analyze these extreme values interactively, providing a deeper understanding of the data and its underlying structure.

This interactive approach, built on the robust RShiny framework, allows for real-time exploration of data, dynamic range selection, and comparison of data subgroups, all of which contribute to the formulation of more accurate and refined hypotheses. Furthermore, the ability to visualize tail behaviors and outliers offers a more comprehensive

view of the data, facilitating the discovery of critical insights that traditional methods often miss. Through its design, *num_dist* not only improves the efficiency of the EDA process but also supports better decision-making and hypothesis generation in various fields, such as finance, healthcare, and environmental science.

As the complexity of data continues to grow, *num_dist* stands as a valuable tool in the data scientist's toolkit. Its ability to reduce reliance on predefined assumptions, enhance exploratory analysis, and provide clear, actionable insights makes it a versatile and essential resource for modern data science workflows. Future developments could further integrate advanced statistical methods like Extreme Value Theory (EVT) to enhance the tool's capabilities, ensuring that num_dist remains at the forefront of EDA innovation.

With its ability to generate hypotheses, detect outliers, and facilitate group comparisons, *num_dist* is a powerful tool for modern data science. It simplifies the process of exploratory data analysis, reduces coding overhead, and provides deeper insights into the most critical areas of a dataset. As the complexity of datasets continues to grow, tools like *num_dist* will become increasingly indispensable in the data scientist's toolkit, enabling more nuanced and impactful analysis.

References

1. Rao, A. S., Vardhan, B. V., & Shaik, H. (2021). Role of exploratory data analysis in data science. In 2021 6th International Conference on Communication and Electronics Systems (ICCES) (pp. 1457 - 1461). IEEE. https://doi.org/10.1109/ICCES51350.2021.9488986
2. Taleb, N. N. (2007). The Black Swan: The Impact of the Highly Improbable. Random House.
3. Sievert, C. (2020). Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC.
4. Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2021). shiny: Web Application Framework for R. R package version 1.7.1.
5. DeBruine, L. M. (2022). Data visualization using R for researchers who do not use R. Advances in Methods and Practices in Psychological Science, 5(2), 25152459221074654.
6. Barnett, D. J., Arts, I. C., & Penders, J. (2021). microViz: an R package for microbiome data visualization and statistics. Journal of Open Source Software, 6(63), 3201.
7. Mowinckel, A. M., & Vidal-Piñeiro, D. (2020). Visualization of brain statistics with R packages ggseg and ggseg3d. Advances in Methods and Practices in Psychological Science, 3(4), 466-483.
8. Gustavsson, E. K., Zhang, D., Reynolds, R. H., Garcia-Ruiz, S., & Ryten, M. (2022). ggtranscript: an R package for the visualization and interpretation of transcript isoforms using ggplot2. Bioinformatics, 38(15), 3844-3846.
9. Singh, G., Singh, J., & Singh, P. K. (2022). Preliminary Analysis of Data Visualization based Uber Technology using R: A Review. In 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS) (pp. 721-725). IEEE.
10. Gu, Z. (2022). Complex heatmap visualization. Imeta, 1(3), e43.