Evaluating the Effects of Visualizing Missing Values on Data Exploration

Hayeong Song* Bahador Saket† John Stasko‡

Georgia Institute of Technology

Abstract

People must often perform analysis on data that contains missing values. We conducted a preliminary empirical study to understand the effect of visualizing those missing values on participants' visual data exploration and decision-making. Study participants purchased a hypothetical portfolio of stocks based on a data set where some stocks had missing attribute values. The stock data was shown in scatter plots. For one group of participants, stocks with missing values simply were not shown, while the second group saw such stocks depicted with estimated values as points with error bars as annotations to indicate missing values and to communicate estimated values. We measured participants' awareness of missing values in the data set and their cognitive load in decision-making. Our results indicate that when missing values were visualized, participants reported a higher awareness level of the missing values, considered a higher number of individual data items, and their decision-making workflow was different.

Index Terms: Information Visualization—Uncertainty Visualization—Decision-Making—Incomplete Dataset;

1 Introduction

Many real-world datasets contain missing values because of various reasons. Designers of data visualizations have faced a dilemma in how to (or how not to) represent those missing values. When creating visualizations, designers have two primary options to manage missing values: 1) simply not show those data items or 2) impute new data values (calculate substitute values) and represent them visually. We know that representing missing data affects people's confidence in their data and results [4, 8]. However, we do not understand as well how representing missing values can impact people's interactive visual data exploration and decision-making processes. More specifically, we hypothesized that:

H1 - Visually representing missing values will generate higher awareness of missing values.
H2 - Visually representing missing values will lead participants to consider more data items.
H3 - Not visually representing missing values will both lead to participants' increased frustration level and mental workload.

These three hypotheses arise from prior work which found that people's awareness of data uncertainty affects how they build trust in data and make decisions [8]. We employ this idea and thus measure participants' level of awareness of missing values and their level of cognitive load involved while performing the task. Prior research in this area has largely focused on how people view and react to the representations of missing data and values in static visualizations [2, 5, 9]. Our work builds upon this prior work by examining how the representation (or lack thereof) of missing values affects people's interactive data analysis process.

*e-mail: hsong300@gatech.edu
†e-mail: saket@gatech.edu
‡e-mail: stasko@cc.gatech.edu

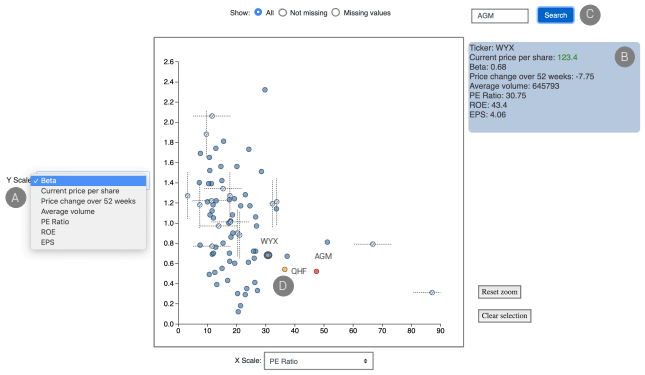

Figure 1: Version of the system that visually represents missing values. Users can (a) select data attributes for the x and y axes using the drop down menus; (b) hover the cursor over a data item (WYX here) to see all of its data attribute values; (c) search for an item by name (result highlighted in red); (d) click on a data item to track it (results colored in orange with labels).

2 User Interface: Interactive Scatterplot

To investigate how people explore and make decisions when a dataset is incomplete, we developed an interactive scatterplot visualization (see Figure 1). Within the scatterplot, each item in the dataset is represented by a small circle. The system supports two versions of the scatterplot that handles the presence of missing values in the dataset differently. The first version (*baseline*) does not visually represent a data item in the view if it is missing a value for either of the two attributes actively being displayed. The second version (*annotation*) visually represents data items that are missing one of the displayed data attributes with error bars. We use a k-nearest neighbors [1] approach with all the other remaining attributes in the dataset as features to find the three most similar items in the dataset. Then we compute the average value of the missing attribute from those three data items. A horizontal error bar line indicates that the x-axis attribute value was estimated and vertical indicates y-axis.

3 Study Design

We conducted a between-subjects experiment to study the influence of representing or not representing missing values on people's decision-making and visual data analysis process. The study contained two conditions: *baseline* and *annotation*. We randomly divided the 18 participants evenly into two groups. Each group worked with one of the conditions.

For the task, participants were given an initial pot of money to spend and build a portfolio of stocks that would return the most profit in the future. We had participants think aloud during the study to help us understand their mental workflow. When each participant finished the task, we asked them to self-report their quantitative ratings of awareness of missing values, mental demand, and frustration in the exit survey to help us understand the effects of visual representation of missing values on cognitive load in decision-

making. These metrics were inspired by the NASA TLX-survey . Scores were reported on a 10-point Likert scale. We also conducted a semi-structured interview to understand participants' decision-making process, the influence of the visualization, user experience, and sought ideas for improving the system. Study materials are available in the supplemental materials.

3.1 Dataset

For the stock buying scenario, we scraped data from Yahoo Finance for 81 different companies and 8 attributes for each stock: ticker symbol, ROE (Return on Equity), EPS (Earnings Per Share), price change over 52 weeks, current price per share, average volume, beta, and PE ratio (Price-to-Earnings ratio). Ticker symbol was a text value while the remaining 7 attributes were numerical values. We anonymized ticker symbols in the dataset to prevent participants from making decisions based on their prior knowledge about a specific company. The initial dataset did not contain missing values, thus we removed at most one attribute per stock. We randomly removed an equal number of low, medium, and high valued attributes.

3.2 Results & Data Analysis

Surveys & Data Item Consideration. In addition to the participants' self-ratings, we tracked the number of data items that the participants "interacted" with or considered. We considered an interacted data item one that participants *tracked* (clicked on) or *searched* by using the capabilities of the system. This metric was inspired by Hindsight design [6], a design that allowed users to be aware of their interaction history that impacted users' data exploration. To analyze data for each metric, we used a T-test. Our results showed support for **H1** & **H2**, that visually representing the missing values led to a significantly increased participants' level of awareness of missing values and data item consideration. However, we did not find support for **H3** as we did not find a significant difference in participants' frustration level and mental demand across two conditions.

Video & Interview Data. We recorded audio/screen during the user study and the follow-up interview. To analyze the video/interview material, we employed Creswell's qualitative data analysis approach [3]. The first author transcribed recordings, coded meaningful text segments, and repeated this process to refine codes, and aggregated similar codes into themes. Some codes had a special focus on participants' ***decision-making workflow***. Our analysis revealed that participants typically followed six steps and we found key differences in some phases across the two conditions.

3.2.1 General Decision-Making Workflow & Findings

In the workflow of their decision-making, participants generally followed six steps.

1. Strategize. Participants developed strategies to make decisions. They determined which attribute values to focus on (e.g., positive price change & beta ≥ 1.0) to help look for appropriate stocks.

2. Explore & 3. Identify Candidates. Participants explored the dataset through the view to gather information. Next, they searched and identified candidate stocks (some with or without missing values) that mostly aligned with their strategy.

4. Weigh the Evidence & 5. Choose Among Candidates. After they identified candidates, participants weighed the evidence to judge if the selected stocks would return future profits. Upon making those judgements, participants selected stocks from the identified candidates for their final portfolio.

6. Review Selections. Lastly, participants reviewed their selections.

Based on our observations, we found key differences across the two conditions in the phases ***strategize***, ***weigh the evidence***, and ***review selections***.

Annotation Condition. When missing values were visually represented, participants' decision-making process was relatively consistent and regular because they were able to ***weigh the evidence*** for data items with missing values. Participants were also more likely to stick with their original ***strategy***. This resulted in participants feeling confident about their decisions when they ***reviewed the selections***.

Baseline Condition. When missing values were not visually represented, participants faced challenges when attempting to ***weigh the evidence*** to select data items for future selection. That is, it was difficult for participants to compare such a data item with other data items. This tended to lead participants to be entangled in that phase and they had to regenerate a new ***strategy***. Additionally, when they ***reviewed the selections***, we did not find comparable evidence that participants felt confident about their decisions.

4 Discussion & Future Work

In this preliminary study, we found important differences between the two conditions. Our results indicated that *visually representing missing values can assist participants to reason about data* and can change their decision-making process. We observed that visually representing missing values helped participants to focus attention on missing values when making decisions. This focus in attention on missing values encouraged participants to make data quality judgments by comparing data items with other items to judge their potential performance. These comparisons and analyses helped participants develop their strategy and decide whether or not to include data items with a missing value, which made participants feel more confident about their selections.

We also found preliminary support for the claim that *not representing missing data* hinders data exploration and analysis. Participants' data exploration and analysis activities were hindered because it was more difficult for them to make judgments about data quality when missing values were not represented, largely due to challenges in comparing items and attribute values. Because exploration and analysis are process-oriented, the repeated challenges created difficulties when participants were trying to make data quality judgments, develop a strategy, and follow it. These differences can be further studied with alternative visual representations of missing data (e.g. statistical values [7]) to observe how different visual representations impact people's behavior and analysis.

References

[1] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[2] R. Andreasson and M. Riveiro. Effects of visualizing missing data: an empirical evaluation. In *2014 18th International Conference on Information Visualisation (IV)*, pp. 132–138. IEEE, 2014.

[3] J. W. Creswell. *Educational research: Planning, conducting, and evaluating quantitative*. Prentice Hall Upper Saddle River, NJ, 2002.

[4] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck. The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(6):697–718, 2003.

[5] C. Eaton, C. Plaisant, and T. Drizd. Visualizing missing data: Graph interpretation user study. *Human-Computer Interaction-INTERACT 2005*, pp. 861–872, 2005.

[6] M. Feng, C. Deng, E. M. Peck, and L. Harrison. Hindsight: Encouraging exploration through direct encoding of personal interaction history. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):351–360, 2016.

[7] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson. When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 5092–5103. ACM, 2016.

[8] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim. The role of uncertainty, awareness, and trust in visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):240–249, 2015.

[9] H. Song and D. A. Szafir. Where's my data? evaluating visualizations with missing data. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):914–924, 2018.