

IBM Research Report

Evaluating the Use of Data Transformation for Information Visualization

Zhen Wen, Michelle X. Zhou
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598



Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Evaluating the Use of Data Transformation for Information Visualization

Zhen Wen Michelle X. Zhou
IBM T. J. Watson Research Center
{zhenwen, mzhou}@us.ibm.com

ABSTRACT

Data transformation, the process of preparing raw data for effective visualization, is one of the key challenges in information visualization. Although researchers have developed many data transformation techniques, there is little empirical study of the general impact of data transformation on visualization. Without such study, it is difficult to systematically decide when and which data transformation techniques are needed. We thus have designed and conducted a two-part empirical study that comprehensively examines how the use of common data transformation techniques impacts visualization quality, which in turn affects user task performance. Our first experiment studies the impact of data transformation on user performance in single-step, typical visual analytic tasks. The second experiment assesses the impact of data transformation in multi-step, complex analytic tasks. Our results quantify the benefits of data transformation in both experiments. More importantly, our analyses reveal that (1) the benefits of data transformation vary significantly by task and by visualization, and (2) the use of data transformation depends on a user’s interaction context. Based on our findings, we present a set of design recommendations that help guide the development and use of data transformation techniques.

Keywords: data transformation, information visualization, smart graphics, automated visualization design, user study

1. INTRODUCTION

In real-world applications, raw data often are not in a “pristine” form that is ready to be visualized and viewed by users. In such cases, the data must be transformed before they can be effectively visualized. For example, raw data may be noisy and need to be cleaned to avoid visual distortion [2]. Inherently complex data may also need to be transformed first (e.g., clustering) to facilitate user comprehension of the resulted visualization [9].

Traditionally, human visualization experts decide when and which data transformations are needed. They manually select and apply the needed transformations to a given data set and then create a visualization of the transformed data [14]. However, such a process becomes impractical in meeting the demands of making information visualization an analytic appliance for the masses [7, 19]. In these cases, average users, who often have limited knowledge of information visualization let alone data transformation, need to use visualization to analyze massive amounts of data. Moreover, many real-world applications such as intelligence analysis require the support of a much more dynamic and flexible process, where both the specific analytic goals and the exact data to be analyzed may be unknown *a priori* [18, 22].

To meet the challenges described above, we envision a smart visualization system, which can assist average users who are not visualization experts in their analytic tasks. Unlike most visualization systems, which know the data to be analyzed and the type of visualization to be used in advance, our envisioned system supports highly dynamic, interactive visual analysis. Given a user’s request (e.g., creating a visual summarization of desired data), the system will dynamically recommend and create visualizations that are tailored to the user’s analytic context at run time.

Due to the highly dynamic nature of user interaction, our

envisioned system cannot anticipate specific data sets to be visualized or whether such a data set is suitable for producing a quality visualization. As a result, the system must *dynamically* determine which data transformations may be needed in response to a user’s request. However, making such a decision is technically challenging, since a number of potentially conflicting factors may affect the decision. For example, data transformation such as sampling helps to reduce data volume and produce a less cluttered visualization. On the other hand, it may result in the loss of information, which in turn affects the fidelity of the resulted visualization [12]. Moreover, some transformation operations are computationally costly (e.g., data clustering). The system must balance the cost and benefit of data transformation, especially in an interactive environment.

To justify and guide our effort in addressing this unique challenge, our first step is to acquire a comprehensive understanding of the impact of data transformation on visualization. To the best of our knowledge, we are unaware of any existing work that can satisfy our needs. We thus have designed and conducted our own study. We hypothesize that data transformation affects visualization quality, which in turn impacts user task performance. Therefore, our study is designed to examine the impact of data transformation on user task performance in various visualization settings. In particular, it answers two sets of questions:

- How much do common data transformation techniques (e.g., data sampling) help users to perform typical visual analytic tasks? Where do we see the significant benefits (e.g., for a particular type of visualization) so we can focus on developing data transformation techniques in these areas to best exploit the benefits?
- How much do common data transformations help users to perform a complex visual analytic task that requires multiple steps? In such a process, whether and what run-time factors, such as a user’s interaction history, influence the choices of data transformation?

To address the first set of questions, we have designed and conducted an experiment to study the impact of data transformation on user performance in a set of typical visual analytic tasks. To answer the second set of questions, we have designed and conducted another experiment to examine the impact of data transformation on user performance in complex, multi-step analytic tasks. We report our observations and analyses in both experiments. We qualitatively and quantitatively assess the impact of data transformation on user task performance under various visualization conditions. We also discuss the implications of our findings to our own work and to the wider information visualization community.

The rest of the paper is organized as follows: we first provide a brief discussion of related work. We then describe our two experiments, including their design, results, and analyses.

2. RELATED WORK

Our work is directly related to research on data transformation for visualization [2, 18]. Researchers have developed a number of data transformation techniques to ensure the creation of effective visualization. Chi has summarized data transformations used for a set of well-known visualizations [5]. To better visualize categorical data, Ma and Hellerstein have developed a clustering-based approach to order nominal data [9]. More recently, transformation techniques such as hierarchical dimension ordering [23] and

ranked low-dimensional projections [15, 21] are introduced to better organize high-dimensional data for visualization. While learning specific transformation techniques from these works, here we systematically investigate the general impact of data transformation on visualization, in turn on user task performance.

In the information visualization community, researchers have also designed and conducted many empirical studies. These studies range from examining specific visualization techniques (e.g., [4, 8, 11, 17]) to evaluating the quality of visualization [10]. There is one empirical study on comparing specific data transformation techniques that prepare high-dimensional data for visualization [16]. However, we are unaware of any empirical studies like ours, which comprehensively examine the impact of data transformation in a wide variety of visualization situations.

3. EXPERIMENT 1

Our first experiment was designed to understand the impact of data transformation on user performance in single-step, typical visual analytic tasks. Our study quantitatively and qualitatively compared user task performance using visualizations of transformed and non-transformed data. The result analyses characterized visualization situations where data transformation is most beneficial.

3.1 Design and Methodology

3.1.1 Experimental system

In this comparative study, we employed the ManyEyes platform for its web accessibility and its support of high-quality visualizations of diverse data sets [19]. For the purpose of our experiment, we augmented ManyEyes to include additional visualizations, such as parallel coordinates [23], to meet the visualization needs required by our study.

We implemented four most commonly used data transformation techniques (Table 1): nominal value ordering [9], dimension ordering [23], data sampling [6], and data cleaning [15]. In addition to controlling the experimental scope, we limited our study to these techniques for four reasons. First, these are general transformation techniques easily applicable to a wide variety of data sets ([2], Chapter 4 in [18]). Second, they represent three distinctive types of typical data operations: (1) data pattern extraction (nominal value ordering and dimension ordering), (2) data volume or complexity reduction (sampling), and (3) data cleaning. Third, it is easy to implement these techniques and tune their parameters to suit a specific data set. Fourth, we can use simple data properties to estimate the usefulness of these techniques for a given data set. For example, if the volume of a data set is large, the sampling tech-

Transformation	Definition
Normalinal value ordering	Order categorical data values by clustering similar data instances
Dimension ordering	Order dimensions by grouping dimensions with high correlations together
Sampling	Uniformly sample data by a given ratio
Cleaning	Separate outliers based on local outlier factor [15]

Table 1. Data transformations used in our study.

nique may be useful. The last criterion is especially relevant to our goal, which is to automate the selection of proper data transformations for a given data set during a human-machine visual dialog.

3.1.2 Tasks

To make our study comprehensive yet manageable, we carefully scope the visualization conditions under which we were to examine the impact of data transformation. Specifically, we characterize each visualization condition along two dimensions: the type of task being performed and the type of visualization being used. As described below, not only can these two dimensions be rigorously defined, but their combinations can also cover a wide variety of typical visualization situations.

Type of user tasks. Existing work has identified a number of user tasks that can be facilitated by visualization [1, 3, 13, 20]. Among these tasks, we decided to focus on three most common types of tasks to control the study scope but still retain the generality of our results. These three types of tasks are: look up, comparison, and distribution. Here are the examples for each type of tasks:

- **Look up:** Name the town with the smallest population.
- **Comparison:** Compare house prices in two cities: New York and Chicago.
- **Distribution:** Characterize the twin birth rate over the last 30 years in the U.S.

Type of visualization. Since we aim to help average users in their tasks, we focused on the use of common visualizations to support the three types of tasks listed above. We adopted four types of visualization from ManyEyes (Figure 1), which accounted for nearly 80% of the usage among the users of ManyEyes [19].

To achieve a balanced, within-subject comparison that covers our entire study space (3 types of tasks, 4 types of visualizations, with or without the use of data transformation), we designed a total of 24 (= 3×4×2) independent tasks. These 24 tasks were divided in 12 (=3×4) pairs, each of which included two similar but not identi-

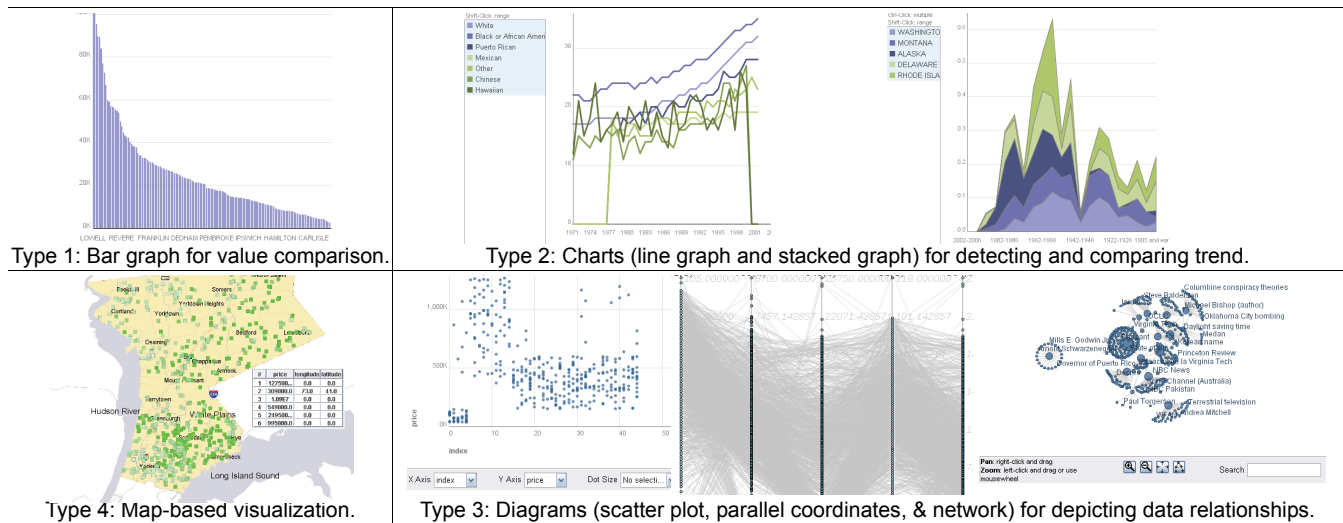


Figure 1. Four types of visualizations used in the Experiment 1.

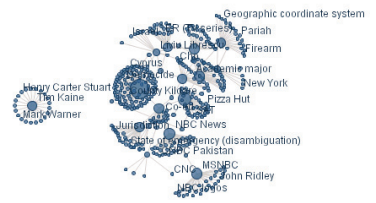

Task	Task A	Task B
Task type	Comparison task	Comparison task
Data set	6151 pairs of co-occurred news topics	27433 pairs of co-authorships for top-1000 authors in PubMed 2005
Visualization	 <p>Type 3 visualization (network diagram)</p>	 <p>Type 3 visualization (network diagram)</p>
Question	Indicate which of the following news topics has <i>most</i> co-occurred topics: A. school shooting, B. China, C. Iraq, D. suicide bombing.	Indicate which of the authors has most co-authors: A. Chen_SC, B. Nakagawa_M, C. Sato_A, D. Kumar_V.
Transformation	Data sampling and clustering	None

Table 2. An example pair of tasks used in Experiment 1.

cal tasks: one to be performed with visualization of transformed data and the other with visualization of non-transformed data. Table 2 shows a pair of such tasks and the complete task list is given in Appendix. For each task, we designed a multiple choice question for a user to answer (Table 2).

For each data set in a task, we used the following guidelines to determine the use of data transformation:

- Use data sampling first if data volume is large.
- Use data cleaning if data appears to be noisy (e.g., the “local outlier factor” is large).
- Use nominal value ordering if similarity metrics can be easily defined to cluster data instances.
- Use dimension ordering if there are more than three data dimensions.

Based on these guidelines, two transformation techniques were applied to the data set of task A mentioned in Table 2.

To avoid biases toward particular data transformation techniques, we made effort to ensure the diversity of data used in our study. For 24 tasks, we chose 24 data sets from 2,100+ data sets contributed by more than 1,400 users on ManyEyes. Among the 24 selected data sets, the number of data dimensions ranged from 2 to 16, and the number of data instances ranged from 30 to 27433. Furthermore, we carefully designed each task to ensure: (1) the task would be best achieved using a visualization, and (2) the visualization used to help achieve the task was indeed the most proper one. These criteria helped to eliminate potential biases toward improper or ineffective use of visualization in our study.

3.1.3 Participants

We recruited twenty participants. There were twelve males and eight females. Their ages varied from mid 20’s to early 50’s. To fit the profile of average users whom we target to help, all our participants were not visualization experts. They all had some experience with general visualization tools (e.g., web-based map tools), but none of them had previous experience with the specific visualization tools used in our experiment.

3.1.4 Methodology

We used a within-subject comparison methodology. We asked each participant to perform all 24 tasks. Among the 24 tasks, 12 would be completed using visualizations of non-transformed data, and the other 12 using visualizations of transformed data. To avoid potential biases like learning effects, we permuted the order and conditions of tasks to cover all combinations.

Before each task, a participant was given a profiling questionnaire to indicate his/her expertise level and frequency of using visualization tools. We gave each participant a 15-minute tutorial, in which we taught the participant how to use the visualization tools required by the experiment.

For each task, the participant was first given the task description and the question to be answered. An interactive visualization was then given for the participant to perform the task. We recorded the task completion time, which was counted from the moment when the participant was given the visualization to the moment when the participant claimed that the task was completed. To prevent a participant from spending too much time on one task, we allotted three minutes per task. If a participant failed to complete a task within the allotted time, we marked the task failed and asked the participant to move on to the next task.

After each task, we collected the participant’s subjective feedback. First, we asked the participant whether the visualization s/he just used had helped to achieve the task. The answers to this question would help us to understand the baseline—whether the visualization helps at all for the given task. Second, we asked the participant to rate the usefulness of the visualization that s/he just used in the task against an alternative visualization on three scales: *better*, *worse*, or *similar*. If the participant had used a visualization of transformed data in the task, the alternative visualization would be the visualization of non-transformed data, or vice versa. We also asked the participant about the rationale for his/her choice.

3.2 Results and Analysis

We quantitatively analyzed the collected data to assess the impact of data transformation on user task performance using both objective (task completion time and task error rate) and subjective measures (participants’ subjective feedback).

3.2.1 Task completion time

We first computed the mean task completion time over all tasks and participants under two settings: with data transformation enabled and disabled. With data transformation enabled, the mean completion time was 40 seconds, while the mean completion time was 89 seconds with data transformation disabled (Figure 2a). The improvement was 55%. A repeated measure variance analysis showed that such a difference was statistically significant: $F(1, 100) = 141, p < 1e-7$.

We then analyzed how various factors might have impacted the task completion time. An ANOVA test found that both task type ($F(2, 100) = 11.4, p < 0.001$) and visualization type ($F(3,$

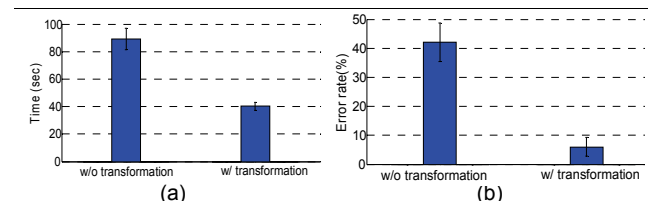


Figure 2. Mean and 95% confidence interval of (a) task completion time, and (b) task error rate with or w/o data transformation.

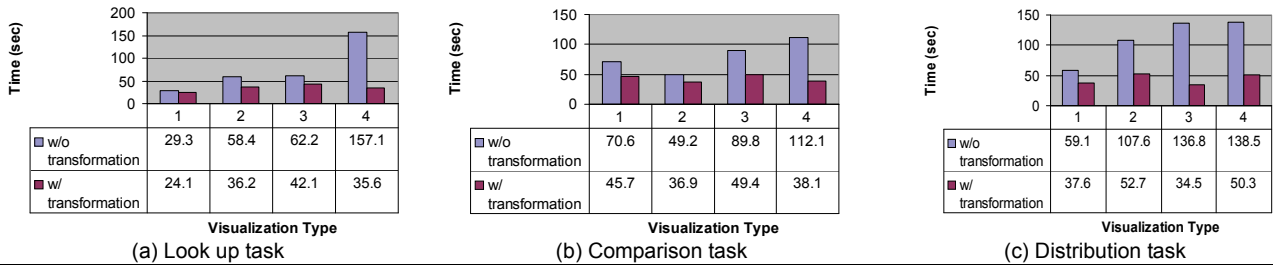


Figure 3. Mean task completion time in various types of tasks and visualizations.

100)=33.1, $p < 1e-7$) had influenced the task completion time significantly (Figure 3). A post-hoc test further identified that the use of data transformation had significantly improved the task completion time in three visualization situations: (1) type 4 visualization (map) for any type of task, (2) type 3 visualization (e.g., network diagram) for comparison and distribution tasks, and (3) type 2 visualization (e.g., line graph) for distribution tasks.

To better understand these results, we further examined the nature of the tasks and the corresponding visualizations used. From our observations, more difficult tasks like identifying data trend or distribution often required users to perform more visual explorations before they could reach their conclusions. Complex visualizations like diagrams (Type 3 in Figure 1) and maps (Type 4 in Figure 1) often encoded complex data sets and required users to spend more time to understand the data. Under these already difficult conditions, should the non-transformed data (e.g., noisy data) further decrease the effectiveness of the intended visualization, a user's task performance would be severely impaired. In such situations, data transformations becomes most beneficial, since it reduces the number of visual explorations needed to complete the task. This helps explain our above results that data transformation is more valuable when complex tasks are to be accomplished or complex visualizations are to be created.

Except the task type and visualization type, our analysis found that other factors had little impact on the completion time. For example, the completion time was not affected by the participants' self-rated expertise level ($F(3, 100) = 0.346, p = 0.79$) or the frequency of using visualization tools ($F(3, 100) = 1.25, p = 0.29$).

3.2.2 Task error rate

Our second task performance metric, *task error rate*, measures the percentage of tasks that the user successfully completed. Our analysis found that the mean error rate across all tasks and participants was reduced from 42.1% (without data transformation) to 6.0% (with data transformation), a 86% reduction (Figure 2b). A repeated measure variance analysis showed that this improvement was statistically significant: $F(1, 100) = 90.1, p < 1e-7$.

Similarly, an ANOVA test found both the task type ($F(2, 100) = 35.5, p < 1e-7$) and visualization type ($F(3, 100) = 13.1, p < 1e-7$) had impacted the error rate significantly. A post-hoc test further identified two visualization situations under which data transformation had helped to reduce the error rate significantly: 1) type 4 visualization (map) for any type of task; and 2) distribution tasks

(Figure 4). Again, these findings suggested that data transformation be more useful for difficult tasks and complex visualizations.

Again, neither the participants' expertise level ($F(3, 100) = 0.387, p = 0.76$) nor the frequency of using visualization ($F(2, 100) = 0.601, p = 0.55$) had any significant impact on the error rate.

3.2.3 Subjective measures

Based on the user feedback collected at the end of each task, we compared the participants' choices between visualizations with and without data transformations.

Our participants overwhelmingly favored visualizations of transformed data, although they were not aware of the fact when they made such a choice. Out of 480 trials (24 tasks \times 20 participants), the participants preferred the visualizations of transformed data in 376 trials (78.5%). Visualizations of non-transformed data were preferred in 23 trials (4.6%). For the remaining 81 trials (16.9%), the participants thought the two visualizations were similar. An ANOVA test showed that task type ($F(2, 100) = 23.0, p < 1e-7$) and visualization type ($F(3, 100) = 6.15, p < 0.001$) were the only two factors that significantly impacted the participants' preferences. Whether the visualization was actually used in the task did not have a significant impact on user preferences ($F(1, 100) = 0.74, p = 0.39$).

To better understand the rationale behind their preferences, we further studied the participants' comments. Based on their comments, users preferred the visualizations of transformed data for two main reasons: (1) visualizations of transformed data were more legible and less cluttered, and (2) visualizations of transformed data directly provided needed information that was otherwise hidden and required much user effort to obtain.

In more than half of the cases (55% of 376 trials) where users preferred the visualizations of transformed data, the corresponding visualizations of non-transformed data were illegible or cluttered. Here is a comment from a participant after using the map tool (Figure 5 a1-a2):

"The map [of non-transformed data] was distorted and had too many occlusions. I couldn't see the town names under the houses. I had to guess the names even after zooming in."

In this case, the distortion caused by noisy data and the visual clutter hid the needed information (i.e., town names).

In the rest of 45% cases, users indicated that visualizations of non-transformed data did not provide the needed information and much user effort was required to obtain the information. The fol-

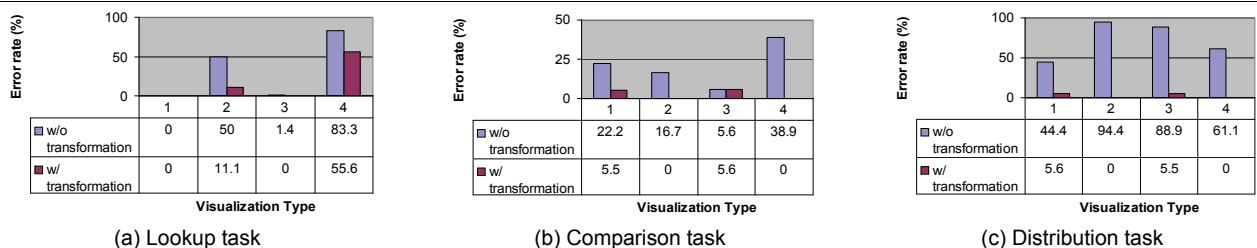


Figure 4. Mean error rate in various types of tasks and visualizations.

lowing comment was on a complex network diagram:

“... this graph [encoding non-transformed data] was too dense to see any distribution. I tried to zoom in but it was very difficult. The system’s response was very slow. It often confused my zoom [interaction] with selection [interaction]. I guess it was because of the overlapping [nodes].”

A similar comment was given on interacting with a parallel coordinates (Figure 5b1–b2):

“The interaction for re-ordering axes was kind of awkward. So I like the alternative plot [of transformed data]. There I could find answers without much interaction.”

The above comments indicated that inherent limitations (e.g., interactivity and performance) in a visualization tool could make a task more difficult. In such cases, data transformation is highly valuable, since a visualization of properly transformed data would reduce user effort to complete the task.

Similarly, our participants appreciated visualizations of transformed data if these visualizations helped them to easily derive data insight. Below is a representative comment on a pair of visualizations like Figure 5 (c1–c2):

“The alternative plot [of transformed data] was more helpful. I could immediately see the pattern of distribution. When I was doing the task [with the plot of non-transformed data], I had to inspect each town one by one to answer the question.”

Although our participants preferred visualizations of transformed data in 78.5% cases, there were nearly a quarter (22.5%) of cases where visualizations of transformed data were not favored. We examined these cases and our findings are two-fold.

First, inherent limitations in a visualization tool could reduce the benefit of data transformation. For example, in our study two line graphs of transformed and non-transformed data were considered similar. The line graph used in our study was very limited and only allowed the use of color instead of position to group lines together. The choices of color were also limited. As a result, it could not take full advantage of the transformed data. Specifically, the transformed data produced an ordered set of lines, which could not be encoded by the line graphs used in the experiment. The comment from a participant confirmed our analysis:

“The positions of the lines were more important to me. The

color grouping didn’t make much difference. The small contrast of the colors did not help either.”

Second, the benefit of data transformation was not evident for very simple visualizations. For example, the bar chart used in our experiment only supported selection but not zooming. One participant commented:

“This bar chart [of transformed data] looked better. But there was only one way [selecting bars] to find the answer. So the two charts were basically the same for the task.”

We also found explanations for cases where visualization of transformed data were considered worse. The main culprit was the loss of information due to transformation. Here is a comment from a participant on a network diagram showing a set of sampled nodes and links:

“The alternative graph [of transformed data] had quite a few isolated nodes. That might be an artifact because the other graph [of non-transformed data] seemed to be a connected graph. Although it didn’t affect my answer, sometimes it may confuse people. There should be a way to let people know some links were dropped.”

Finally, we checked participants’ comments for potential biases toward the possible misuse of visualizations for designated tasks. Only in 4 trials out of 480 trials (0.83%), two participants commented that other visualizations or alternative data presentations like data tables might better help to accomplish the tasks. Therefore, we felt confident that we had used the proper visualizations for most of the tasks.

3.3 Results Summary and Recommendations

Our first experiment demonstrates that standard data transformations significantly improve users task performance, especially in helping reduce their task completion time (55%) and task error rate (86%). Among various factors, the task type and visualization type are the only two factors that have significant impact on the usefulness of data transformation. Specifically, data transformation is more valuable to users when they perform difficult tasks or use complex visualizations. Users’ subjective ratings and comments concurred with our findings.

Based on our findings, we make the following recommenda-

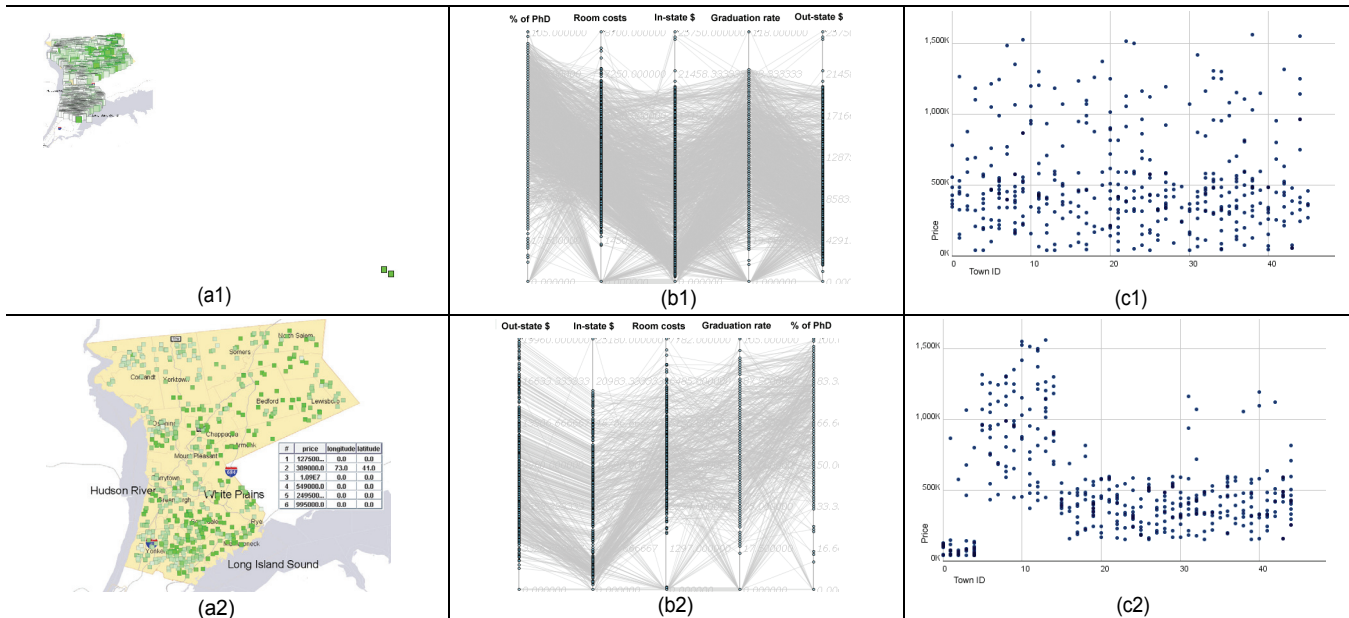


Figure 5. Visualizations for 3 tasks. Display (a1) shows a set of houses on map; Display (b1) shows five attributes of a set of colleges; Display (c1) shows house prices (Y axis) in a set of towns (X axis). Displays (a1), (b1), (c1) are created using non-transformed data. Displays (a2) is created using data after cleaning. Display (b2) is created using data after sampling and dimension ordering. Display (c2) is created using data after nominal value ordering.

tions that may be used to guide the development and application of data transformation techniques.

Types of transformation techniques. Our result analyses show that different transformation techniques have different effects on a visualization. Based on the data transformation effects observed in the study, we abstract three types of transformation techniques:

- **Regulatory techniques:** clean and normalize data (e.g., outlier separation technique).
- **Scaling techniques:** adjust data volume and complexity (e.g., sampling technique).
- **Organizational techniques:** identify inherent structures of data and organize the data accordingly (e.g., nominal value ordering technique).

Guidelines of applying data transformation. Our results quantitatively identify that data transformation is more beneficial when users perform difficult tasks (e.g., distribution tasks), or use complex visualizations (e.g., complex diagrams like parallel coordinates). To best exploit the benefits of data transformation, we should probably target developing transformation techniques supporting these tasks and visualizations first. Based on the identified characteristics of these tasks and visualizations, we derive a set of usage guidelines for data transformation. These guidelines are based on the ones used in our study but generalized to best exploit the benefits of various types of data transformation:

- Use regulatory techniques if data is noisy.
- Use scaling techniques (e.g., sampling) if data volume is large and the task is *not* for fine-grained data analysis.
- Use organizational techniques (e.g., clustering) to extract inherent structures in data, if the intended visualization is complex (e.g., parallel coordinates) or the task is to make data comparisons or examine data distribution/trend.
- Do *not* use data transformation if the target visualization cannot render the properties of transformed data.

Visualization quality metrics for measuring the benefit of data transformation. Our analyses of participants' visualization preferences suggest how humans judge the quality of a visualization, in turn the benefit of data transformation. Accordingly, we extract three features of measuring visualization quality:

- **Visual legibility.** An effective visualization must be legible. Data transformation is beneficial, if it helps to produce a legible visualization.
- **Visual pattern recognizability.** An effective visualization assists users in detecting visual patterns to gain insights. Users favor visualizations of transformed data (e.g., dimension ordering) that allow them to more easily detect patterns.
- **Visual fidelity.** Data transformation may alter data properties. If not careful, the interpretation of data may be subverted. For example, a connected graph may appear disconnected after data sampling. To avoid misinterpretations, data transformation should always try to maintain the truthfulness of the original data.

These abstracted features provide a starting point for us to develop computational metrics that formalize the benefit of data transformation. This is a critical step toward our goal, which is to let a system dynamically decide whether and what data transformation is needed for a given situation. To do so, the system must be able to computationally evaluate and compare the visualization quality before and after the use of data transformation. Such information helps the system to balance the cost and benefit of using a particular data transformation (e.g., data sampling helps improve visual legibility but may reduce visual fidelity).

Interactive data transformation. As indicated in our study, data transformation may not always be desirable. To ensure user satis-

faction, we also recommend that the system let users be aware of the transformation used and allow them to control the transformation if needed. If the users do not like the system-chosen transformation, they can always change or disable it.

4. EXPERIMENT 2

Our first experiment examined the impact of data transformation on user performance in single-step, typical visual analytic tasks. Our second experiment was designed to study the impact of data transformation in multi-step, complex analytic tasks. In particular, we would like to study how a user's evolving context impacts the choice and benefit of data transformation.

4.1 Design and Methodology

4.1.1 Experimental system

For the purpose of this experiment, we built a Wizard-of-OZ (WOZ) system that let a human moderator to respond to a participant's analytic request. Given a request, the moderator dynamically chose the desired data set, the type of visualization, and the needed data transformations. The WOZ system would then create an interactive visualization for the given data. It supported all four types of data transformations and a sub-set of visualizations used in Experiment 1.

To achieve fast response time, our WOZ system let the moderator use a GUI to quickly specify a data set, visualization type, and data transformations. For example, given a user request "*houses under \$600K*", the moderator would (1) enter a data constraint in a GUI constraint box *price* < \$600K, (2) select the map visualization from a check list of available visualizations, and (3) select a data cleaning transformation from a check list of available transformation techniques. The moderator could also customize the parameters of a selected transformation technique in the GUI. To assist the selection of data transformations, we implemented procedures that automatically extract data properties of a given data set (e.g., data volume and outlier factor). During the study, the moderator would use the extracted data properties and the chosen type of visualization to select suitable data transformations.

4.1.2 Tasks

In this experiment, we studied the impact of data transformation on user performance in exploratory tasks. In such tasks, users were asked to explore the data space and find one or more data targets that meet a set of fuzzy criteria (e.g., cheap houses in good school districts). We chose to focus on multi-criteria, exploratory search tasks for two reasons. First, users often need to take multiple steps to explore the data space and balance the criteria. Thus, a user's interaction context plays an important role in such tasks. In this experiment, such behavior would help us study whether and how the context may influence the choice of data transformation. Second, exploratory search still allows us to objectively compare user performance (e.g., task completion time and error rate).

In this experiment, we designed two similar but not identical exploratory search tasks in the real estate domain. In one task, the criteria to be traded off for target houses were: 1) large, 2) cheap, and 3) in good school districts. In the other task, the criteria were: 1) new, 2) expensive, and 3) close to water. We used a realistic data set obtained from multiple listing services, which included 2000+ houses, 70+ towns, and 400+ schools. The data set covered 25+ distinctive data concepts (e.g., house and city), each with a number of attributes (e.g., a house has 40 attributes, and a city has 25).

4.1.3 Participants and Methodology

The participants in our first experiment were invited in this experiment. Similarly, we employed a within-subject methodology. Each participant was asked to perform both tasks described above, one with data transformation enabled and the other disabled. We allotted 15 minutes for each task and permuted the task order to cover

all combinations.

In each task, a participant was first given a task description. S/he could then make a request to evaluate any of the three criteria. However, s/he must use concrete constraints to evaluate a fuzzy criterion, e.g., using “*price < \$600K*” for the “*cheap*” criterion. The participant could evaluate new criteria in the context of or independent of his/her previous exploration. For example, s/he might evaluate the “*large*” criterion in the context of “*cheap*” houses found so far. The participant could also interact with the generated visualization to examine the retrieved data.

Each task ended when the participant claimed to have completed the task or the allotted time was up. The participant was asked to record the data target(s) found and the rationale substantiating their findings.

4.2 Results and Analysis

Like in our first experiment, we used the task completion time and error rate to quantitatively analyze the impact of data transformation on user task performance. We then examined how run-time factors, such as a user’s interaction history, had impacted the use of data transformations over the course of a task.

4.2.1 Task completion time

In addition to computing the overall task completion time, we analyzed the time spent at each step of a task. We timed a step from the moment when a participant was shown a visualization to the moment when the participant made his/her next request¹. Steps were also indexed by the order as they occurred. These measures let us study the impact of data transformation on different stages of a task. An ANOVA test showed that the use of data transformation ($F(1, 92)=55.2, p<1e-4$) and the index of a step ($F(5, 92)=26.7, p<1e-7$) had significantly impacted the time spent at each step. Neither the search criterion ($F(3, 92)=0.71, p=0.49$) nor the task order ($F(1, 92)=0.021, p=0.89$) had any significant impact.

Furthermore, we performed a linear regression analysis between two variables: the time spent at a step and the index of a step. We examined their relationships in two conditions: with or without data transformation. The results indicated that data transformation helped to reduce the task time more at earlier steps of a task (Figure 6a). Our observations and the participants’ comments provided us explanations to this phenomenon. At the beginning of a task, the participants were exploring the data space to get an understanding of the space. In these steps, the participants typically worked with larger and complex data sets. Consistent with our findings in Experiment 1, data transformation greatly helped users in these difficult situations.

We also studied the impact of data transformation on the total number of steps needed to complete the tasks. An ANOVA test with a post-hoc test (Figure 6b) showed that the use of data transformation was the only factor that had significant impact on the

number of steps taken: $F(1, 32) = 17.0, p < 0.001$; and it helped to reduce the number of steps (mean value) significantly from 4.60 (without transformation) to 3.70 steps (with transformation).

Finally, we examined the overall task completion time. An ANOVA test showed the use of data transformation was the only factor that significantly affected the overall completion time: $F(1, 32) = 17.0, p < 0.001$. A post-hoc test also showed that the mean task completion time was significantly reduced from 767 seconds (without data transformation) to 432 seconds (with data transformation), amounting to 43.7% reduction (Figure 6c).

4.2.2 Task error rate

We asked two domain experts who had extensive real estate knowledge and were familiar with the data set to score the targets found by the participants. The score was between 0 and 1 depending on how well the found targets satisfied the given criteria. An ANOVA test with post-hoc analysis showed that mean value of the error rate was reduced from 11.0% (without data transformation) to 3.5% (with data transformation). However, the difference was not statistically significant (Figure 6d). Compared to Experiment 1, the benefit was also greatly reduced. Based on our observations, we attributed the reduced benefit to two main reasons. First, in Experiment 1, the participants had only one shot (one visualization) to get the needed information. In contrast, in Experiment 2, the participants could get needed information from multiple steps. For example, a participant might not be able to identify certain town names on a map at one step, but s/he could see the names at other steps. Second, in Experiment 2 the participants could re-evaluate previous criteria if they later found they had made mistakes. But they could not do so in Experiment 1. In general, regardless the use of data transformation, Experiment 2 was more forgiving in terms of affording users’ mistakes during their tasks.

4.2.3 Impact of run-time factors on data transformation

Our second experiment was also designed to examine whether and which run-time factors might influence the choices of data transformation used over the course of a task. Here, a choice of data transformation (d_i), which included one or more transformation techniques like sampling and cleaning, was made at a user step (s_j). In addition, there were three run-time factors associated with a step: (1) the search criteria used (q_i), (2) the type of visualization used (v_i), and (3) a sequence of user steps leading to this step (seq_i)—a user’s interaction history.

To quantitatively analyze the relationships between the three run-time factors (v_i, q_i, seq_i) and the choices of data transformations (d_i), we “normalized” the collected data. We collected a total of 74 sets of data transformation used, representing seven *distinct* choices. We also clustered all 74 sequences (seq_i) into six groups based on their similarity. The similarity between two sequences were measured from three aspects: (1) the query used at each step, (2) the properties of the query results (e.g., data volume [6], outlier

¹ We excluded steps for timing where empty visualization was generated due to the empty data retrieved for a user’s request.

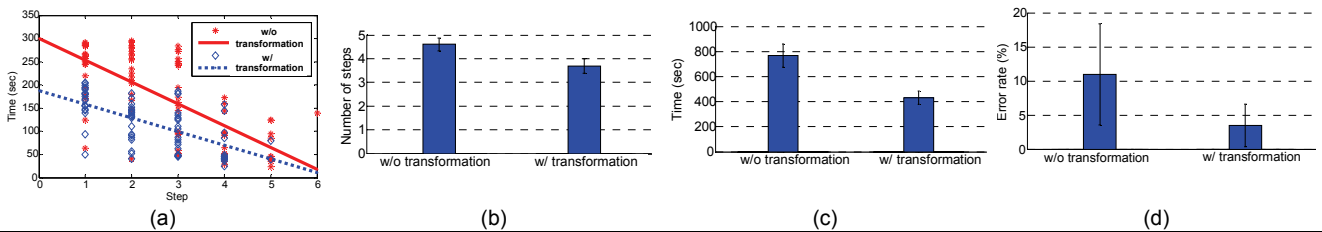


Figure 6. a) Time spent at a step. The red solid line is a linear regression for conditions with non-transformed data, and the blue dotted line is for conditions with transformed data. b) Mean and 95% confidence interval of number of steps to complete tasks with and without dynamic data transformation. c) Mean and 95% confidence interval of completion time with and without dynamic data transformation. d) Mean and 95% confidence interval of error rate with and without dynamic data transformation.

factor [15]), and (3) the type of visualization used at each step.

An ANOVA test showed that *all* three run-time factors significantly impacted the choices of data transformation: the user's interaction history ($F(5, 67) = 12.2, p < 1e-4$), the search criteria ($F(2, 67) = 15.2, p < 1e-4$), and the type of visualization ($F(4, 67) = 6.40, p < 0.001$). We then performed a linear regression test to understand the contributions of each of the three factors to the choices of data transformation. Our regression result showed that the user's interaction history accounted for the biggest portion of the variances in the data transformation (32% vs. less than 1% for other two factors). This confirmed our hypothesis that a user's interaction context, especially the interaction history, influences the choices of data transformation. As a result, data transformations must be derived dynamically in a user's analytic context.

4.3 Result Summary and Recommendations

Like Experiment 1, our second experiment verified that common data transformations significantly improved user performance in complex tasks. Moreover, it showed that a user's interaction context significantly influenced the choice of data transformation.

Based on these findings, we form two recommendations to guide the use of data transformation in support of complex analysis tasks. First, the choice of data transformation should be evaluated at every interaction step, especially at earlier stages of a long task. As shown in our study, data transformation helps improve user task performance (task completion time) significantly in earlier stages of a task. Second, we could leverage the state information (e.g., data exploratory stage versus data lookup stage) to guide the selection of data transformation. For example, data scaling may be useful (e.g., sampling) to provide users with an overview of the data space at an exploratory stage, but may not be desirable when users intend to make fine-grained data comparison.

5. CONCLUSIONS

Data transformation, like data cleaning and sampling, prepares raw data for effective visualization. In this paper, we present two comparative experiments on examining how data transformation impacts user task performance in various visualization situations (e.g., varied tasks and the types of visualization used). Our analyses show that data transformation significantly improves user performance in both single-step and multi-step analytic tasks. The benefit of data transformation is most significant in difficult tasks or with the use of complex visualization tools. Our experiments also conclude that a user's interaction context dictates the choices of data transformation. Therefore, data transformations must be derived dynamically based on a user's interaction context.

Based on our findings, we have made several recommendations to guide the development and use of data transformation to best exploit its benefit. Specifically, we have identified three types of data transformation techniques that help to produce a quality visualization. We have also abstracted a set of guidelines that suggests when and what types of data transformation are most useful. To systematically assess the benefit of data transformation, we have extracted three visualization quality measures, the result indicators of data transformation. Finally, we have suggested that a user's interaction context, such as the state of a task, be used to help decide the need and type of data transformation in context.

REFERENCES

[1] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *InfoVis '05*, pages 111–117, 2005.

[2] S. Card, J. Mackinlay, and B. Shneiderman, editors. *Readings in In-*

formation Visualization. Morgan Kaufmann, 1999.

- [3] S. Casner. Task-analytic approach to the automated design of graphic presentations. *ACM Trans. Graphics*, 10(2):111–151, 1991.
- [4] C. Chen and M. Czerwinski. Empirical evaluation of information visualization: An introduction. *Int'l J. Human Computer Studies*, 53(5):631–635, 2000.
- [5] E. Chi. A taxonomy of visualization techniques using the data state reference model. In *InfoVis '00*, pages 69–76, 2000.
- [6] Q. Cui, M. Ward, E. Rundensteiner, and J. Yang. Measuring data abstraction quality in multiresolution visualization. *IEEE Trans. on Vis. and Comp. Graphics*, 12(5):709–716, 2006.
- [7] C. Johnson, R. Moorhead, T. Munzner, H. Pfister, P. Rheingans, and T. Yoo. Nih/nsf visualization research challenges. <http://www.sci.utah.edu/vrc2005/vrc-report-vis05.pdf>.
- [8] A. Kobsa. User experiments with tree visualization systems. In *InfoVis '04*, pages 9–16, 2004.
- [9] S. Ma and J. Hellerstein. Ordering categorical data to improve visualization. In *InfoVis '99*, pages 15–18, 1999.
- [10] C. North. Visualization viewpoints: Toward measuring visualization insight. *IEEE Computer Graphics & Applications*, 26(3):6–9, 2006.
- [11] K. Ridsen, M. Czerwinski, T. Munzner, and D. Cook. An initial examination of ease of use for 2d and 3d information visualizations of web content. *Int'l J. Human Computer Studies*, 53(5):695–714, 2000.
- [12] B. Rogowitz and L. Treinish. How not to lie with visualization. *Computers in Physics*, (3):268–274, 1996.
- [13] S. Roth and J. Mattis. Data characterization for intelligent graphics presentation. In *CHI '90*, pages 193–200, 1990.
- [14] P. Saraiya, C. North, V. Lam, and K. Duca. An insight-based longitudinal study of visual analytics. *IEEE Trans. on Vis. and Comp. Graphics*, 12(6):1511–1522, 2006.
- [15] J. Seo and B. Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization*, 4(2):96–113, 2005.
- [16] J. Seo and B. Shneiderman. Knowledge discovery in high-dimensional data: Case studies and a user survey for the rank-by-feature framework. *IEEE Trans. on Vis. and Comp. Graphics*, 12(3):311–322, 2006.
- [17] J. Stasko, R. Catrambone, M. Guzdial, and K. McDonald. An evaluation of space-filling information visualizations for depicting hierarchical structures. *Int'l J. Human Computer Studies*, 53(5):663–694, 2000.
- [18] J. Thomas and K. Cook, editors. *Illuminating the Path: The R&D Agenda for Visual Analytics*. IEEE, 2005.
- [19] F. Viegas, M. Wattenberg, F. V. Ham, J. Kriss, and M. McKeon. Many eyes: A site for visualization at internet scale. In *InfoVis '07*, 2007.
- [20] S. Wehrend and C. Lewis. A problem-oriented classification of visualization techniques. In *IEEE Vis '90*, pages 139–143, 1990.
- [21] L. Wilkinson, A. Anand, and R. Grossman. High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. *IEEE Trans. on Vis. and Comp. Graphics*, 12(6):1363–1372, 2006.
- [22] W. Wright, D. Schroh, P. Proulx, A. Skaburskis, and B. Cort. The sandbox for analysis: concepts and evaluation. In *CHI '06*, pages 801–810. ACM.
- [23] J. Yang, W. Peng, M. Ward, and E. Rundensteiner. Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *InfoVis '03*, pages 105–112, 2003.

APPENDIX

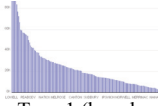

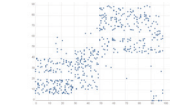


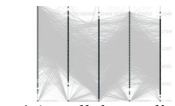
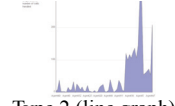

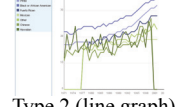

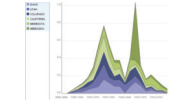
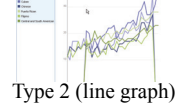
Task ID	Task type	Data set	Visualization	Question	Transformation
A1	Look up	Population of 150 towns in US	 Type 1 (bar chart)	Find the town with the smallest population.	Cleaning
A2	Compare	A set of houses in northern Westchester, NY	 Type 3 (map)	Use the visualization to compare house tax in Crompond and Mt. Kisco.	None
A3	Distribution	Time and server name of 512 events occurring on 100 computer servers.	 Type 4 (scatter plot)	From the visualization, mark which of the following is true: A. About (1/2) of the servers have events occurring in time range [0, 50]. B. About (2/3) of the servers have events occurring in time range [0, 50]. C. About (1/3) of the servers have events occurring in time range [0, 50]. D. None of the above. D. Can not tell.	Normal value ordering
A4	Look up	A set of houses in southern Westchester, NY	 Type 3 (map)	Name the town that is close to Long Island Sound.	None
A5	Compare	6151 co-occurred news topics	 Type 4 (network diagram)	Indicate which of the following news topics has <i>most</i> co-occurred topics: A. school shooting, B. China, C. Iraq, D. suicide bombing.	Sampling & clustering
A6	Distribution	6 attributes of 1304 colleges	 Type 4 (parallel coordinates)	Identify two axes that have the most distinct correlation.	None
A7	Look up	Number of calls handled by 57 agents	 Type 2 (line graph)	Use visualization to identify the agents whose number of calls handled is between 100 and 120	Normal value ordering
A8	Compare	6151 co-occurred news topics	 Type 4 (network diagram)	Indicate which of the following news topics has <i>most</i> co-occurred topics: A. mass murder, B. the U. S. secret service, C. Pakistan, D. firearm.	None
A9	Distribution	Twin birth rate of seven ethnic groups from 1974-2004	 Type 2 (line graph)	Characterize the twin birth rate over the last 30 years in US and access which of the following is true: A. The trend of Chinese is most different from others. B. The trends of Black and White are most similar. C. The rate of Mexican has the largest increase. D. The trend of Black is most different from others.	Dimension ordering
A10	Look up	A set of houses in Westchester, NY	 Type 3 (map)	Use the visualization to identify how many towns are along the Hudson River	None
A11	Compare	# of bridge problems in 6 US states	 Type 2 (stack graph)	Name two states with similar bridge problem trend	Dimension ordering
A12	Distribution	Twin birth rate of six groups in last 10 years	 Type 2 (line graph)	Identify the ethnic group that has the most distinct trend of twin birth rate change	None

Table 3. Task A1-A12 in Experiment 1

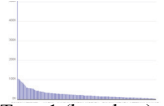

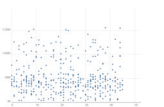



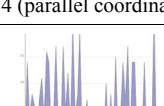
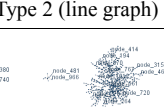
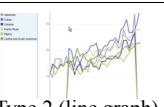
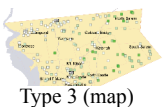


Task ID	Task type	Data set	Visualization	Question	Transformation
B1	Look up	Population densities of 150 towns	 Type 1 (bar chart)	Identify the town with the smallest population density.	None
B2	Compare	A set of houses in Westchester	 Type 3 (map)	Compare house land in Bronxville and Harrison and find in which town houses have larger land.	Cleaning & sampling
B3	Distribution	Prices of houses in 50 towns	 Type 4 (scatter plot)	Use the visualization to assess which of following is true: A. About (1/2) of the towns have houses below \$500K. B. About (2/3) of towns have houses below \$500K. C. About (1/3) of towns have houses below \$500K. D. None of above.	None
B4	Look up	A set of houses in Westchester	 Type 3 (map)	Name the town that is close to the norther county border.	Sampling
B5	Compare	27433 pairs of co-authorships for top-1000 authors in PubMed 2005	 Type 4 (network diagram)	Indicate which of the authors has most co-authors: A. Chen_SC, B. Nakagawa_M, C. Sato_A, D. Kumar_V.	None
B6	Distribution	4 attributes of 2400 houses	 Type 4 (parallel coordinates)	Using the diagram to identify which two of the house attributes have the most distinct correlation.	Dimension ordering
B7	Look up	Average idle time of 57 call center agents	 Type 2 (line graph)	Identify the number of agents whose average idle time 10 and 15.	None
B8	Compare	A network of 3112 computers	 Type 4 (network diagram)	Indicate which of the computer has most connections: A. node_194, B. node_414, C. node_740, D. node_767.	Sampling
B9	Distribution	Twin birth rate of six groups in last 10 years	 Type 2 (line graph)	Identify the ethnic group that has the most distinct trend of twin birth rate change	None
B10	Look up	A set of towns in Westchester	 Type 3 (map)	Identify the number of towns near the north boundary of Westchester.	Sampling
B11	Compare	# of bridge problems in 8 US states	 Type 2 (stack graph)	Identify two states with similar bridge problem trend.	None
B12	Distribution	27433 pairs of co-authorships for top-1000 authors in PubMed 2005	 Type 4 (network diagram)	Find the biggest author cluster that has most members.	Sampling

Table 4. Task B1-B12 in Experiment 1