Comparing the effectiveness of visualizations of different data distributions

Ariane M. B. Rodrigues, Gabriel D. J. Barbosa, Hélio Lopes, and Simone D. J. Barbosa
Department of Informatics, PUC-Rio, Rua Marques de Sao Vicente, 225, Gavea, Rio de Janeiro, 22451-900, Brasil
arodrigues@inf.puc-rio.br, gabrieldjb@gmail.com, lopes@inf.puc-rio.br, simone@inf.puc-rio.br

***Abstract*—The effectiveness of a data visualization depends on several factors, such as the visual encoding of its elements, the analytical task it aims to support, the dimensionality of the data, and even the data distribution. In this work, we report an empirical evaluation of common data visualizations, focusing on how different data distributions affect their effectiveness and the level of confidence users have when answering questions related to certain analytical task types. The study allowed us to assess to what extent some data visualizations are more effective than others, regardless of data distribution. We conclude that data visualization creators and algorithms need to consider the data distribution when generating data visualizations. From the results, we also propose recommendations for choosing a visualization when dealing with data distribution issues.**

I. Introduction

Data visualization is an important tool for exploring, analyzing, and presenting data characteristics, from the most obvious to the least evident. It characterizes a common language, which transforms raw data in something that can be interpreted and communicated [1]. In fact, some data characteristics may be easier to identify and analyze through visual representations, for example, detect patterns.

There has been a great effort in defining visualization recommending systems that suggest better visualizations based on certain concepts, such as data characteristics [2]–[12]. Likewise, much related work has sought to evaluate the effectiveness of diverse visualizations [8], [13]–[23]. The former focus on design: the combination of data and task should determine the best visualization; whilst the latter focus on evaluation: the combination of data and visualization allows assessing whether and to what extent the task was successful [24].

Visualization tasks may be defined as the goals that analysts have when visualizing data. Since the 90s, there has been interest in identifying and classifying such tasks [5], [25], in a way so as to map them onto efficient visualizations and, if possible, to do so automatically. Later, there has been greater interest in formalizing these classifications in taxonomies for diverse ends within the data visualization field [6], [14], [26]–[32]: lists of verbal task descriptions, mathematical task models, domain specific task collections, and combinations of different procedural tasks in workflows [24].

In this context, we can consider that a visualization allows answering certain analysis questions on some data. Such questions can be classified according to a visualization task taxonomy [33]. Our main goal in this work is to identify how well some common visualization types support an analyst in answering certain analysis questions given a data set. Despite some related work already discussing this topic [34], [35], they do not take into consideration the data distribution. Data distribution may affect data visualization effectiveness and efficiency, especially when there are very close similarities or some large discrepancies within the values. In a bar chart, for example, if a bar is much longer than others, reading and comparing may be difficult due to scaling issues. Figure 1 exemplifies this problem: in chart A, it is possible to quickly identify which year had the lowest value (2005); however, for chart B, the identification becomes more difficult (July). To reach our goal, we conducted an empirical study where we identified to what extent the data distribution affects its effectiveness, given an analytical question.

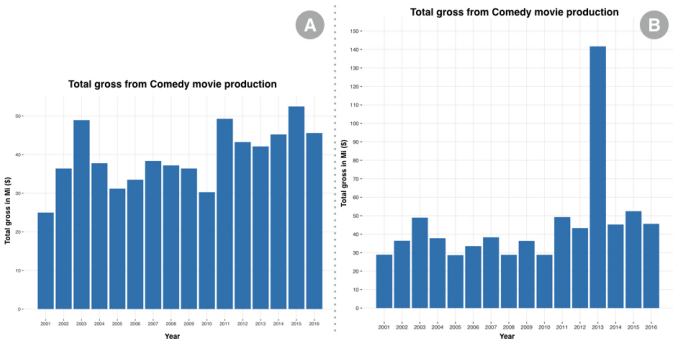

Figure 1. Bar chart representing different data distributions, making it easier (A) or harder (B) to identify certain values.

The remainder of this paper is organized as follows. Section II presents existing work on information visualization. Next, we report an empirical study that evaluated the visualization effectiveness and efficiency in answering certain data analysis questions (section III) and its results (section IV). Finally, we discuss the study's threats to validity (section VI) and point to some relevant future work (section VII).

II. Related Work

This section presents two groups of related work: taxonomies of visualization tasks and empirical studies to evaluate the effectiveness of different types of visualization.

A. Taxonomies of Visualization Tasks

Visualization tasks have been an object of study since the 1990's. Wehrend and Lewis [25] defined a classification scheme that maps objects (data attributes) and "operations"

(representation objectives) to find an appropriate visualization technique for a given problem – the user's goal in analyzing the representation. Roth and Mattis [5] classified visualization problems and their solutions independently of domain, and proposed a taxonomy of information characteristics which provides a list of different user objectives in seeing a visual representation. Their proposed classification is very similar to Wehrend and Lewis', albeit more succinct and focused on the automatic generation of a representation. These taxonomies are considered low-level and user focused.

Shneiderman [6] proposed TTT (Task by data type), a high-level, system-focused taxonomy based on data types and on the problem the user seeks to solve. He wanted to guide graphical user interface design for data visualization analysis.

Amar et al. [33] defined 10 low-level analysis tasks that a person may perform when working with data. They defined "aggregate functions", which create a numeric representation for a set of entities in the data set. They claim that high-level tasks do not express a specific objective or task, but require an answer for a more direct question, which is usually derived by using one or more low-level analytic operations [36]. Some of these questions may be answered by text in an efficient way; others require visualizations for an efficient answer. However, even when a textual representation is considered sufficient to answer a certain question, visualization may amplify the understanding of an answer and its context. Later studies became more specific, as is the case of Lee et al. [30], which defined the list of chart visualization tasks with enough detail so that it would be useful both for designers who seek to improve their systems and for evaluators who seek to compare chart visualization systems. In contrast, all tasks were composed of tasks created by the primitive tasks described by Amar et al. [33], as well as two generic tasks and one chart-specific task. Chen et al. [37] explored tasks related to "data, visualization and objective", and defined a taxonomy to categorize facts that may be extracted from multidimensional data, in a visual data analysis task. Facts are patterns, relationships or anomalies extracted from data through analysis [37].

More recently, Brehmer and Munzner [31] asserted that visualization tasks ought to be described in an abstract fashion, through different levels: *why* the task is conducted, *how* the task is conducted and *what* are the task inputs and outputs. The low-level taxonomies proposed previously only answer *how*, while the high-level ones only answer *why*. For this reason, some researchers consider the use of more than one taxonomy, as is the case of VLAT [20]. To develop VLAT, a visualization literacy test, Lee et al. associated tasks from three different taxonomies. First, they combined the low-level taxonomy [33] with the facts-based one [37], and afterwards discarded some of the tasks that were included in *how* and *why* from [31] – the discarded tasks were related to manipulation and generation of new elements, and not reading and interpretation of visual representations of data. VLAT is an especially relevant work to our own. It proposes metrics for difficulty indices and discrimination in visualization evaluations, which served as a basis for comparison in our study. We have chosen the 12 tasks resulting from the association in their proposed taxonomy, as explained in Section III.

B. Evaluating effectiveness of different types of visualization

Many empirical studies have evaluated the effectiveness of different types of visualization. Some are specific to the chart type: bar charts [36], [38], scatterplots [21], [35] and time series [39], [40], for example. Others compare two types of visualization: bar *vs.* line charts [41], tables *vs.* pie charts [42] and bar *vs.* radar charts [43]. These studies were conducted under different combinations of data sets and tasks.

Saket et al. [14] used crowdsourcing to evaluate the effectiveness (*i.e.*, proportion of correct answers) of five types of bi-dimensional visualizations in small scale (5-34 data points) for two data sets (cars and films). They evaluated tables and line charts, bar charts, scatter plots and pie charts. They chose few data points because, for more than 50 data points, they would face two challenges: difficulty in task completion, and task duration over 2 minutes. However, some of their conclusions cannot be generalized to a larger dataset, with more categories. For example, the pie chart was one of the most effective charts for finding an extreme value (minimum or maximum value). For a data set with many categories, this same chart would have too many slices, perhaps some very similar ones, decreasing their effectiveness for this task.

Lee et al. [20] developed a Visualization Literacy Assessment Test (VLAT), using a systematic approach based on Psychological and Educational metrics. They aimed to evaluate the ability to read and interpret data that is visually represented and to extract information from them. To generate the final set of test items, they went through six stages, three of which involving user studies, as well as a final stage to evaluate visualization literacy and the skill to learn non-familiar visualizations through a comprehension test. VLAT contains 53 items that combine: visualization $\times$ visualization task $\times$ question $\times$ CVR (measures how essential a task is for visualization literacy) $\times$ P (task difficulty) $\times$ D (discrimination). We used VLAT to select the task-chart type combinations we would use.

Kim and Heer [21] conducted a study to evaluate performance at different task types (comparison of individual values *vs.* aggregate values) and data distribution (cardinality and entropies). They used four types of analysis tasks and five variations of visual codings (alternating the analysis variable in x, y, color, size, and position). The data sets had at most 30 points and only 3 analysis variables were used (1 categorical and 2 quantitative). All of the evaluated visualizations were chart variations at the position of points (in a scatterplot) or at the mapping of some variable onto the point size (bubble). The questions had only two available answers, reducing the opportunities for errors. Despite having balanced the number of people in the task and data distributions, each person always answered the same task and distribution. There were 8 different questions for each coding. Our work extends Kim and Heer's in various aspects: (i) we consider more than one visualization type; (ii) the data set used in our study contains

a significant number of data points (3,722), making it more realistic; (iii) we consider a wider range of visualization tasks; and (iv) we compare our results to other studies through predefined metrics. In this work, we seek to identify, for each task, the best type of visualization, in terms of effectiveness, time on task, and adequacy to the task.

III. Survey Study Plan

We conducted a survey study to investigate the performance (effectiveness and efficiency) of different chart types for given visualization tasks, given certain data distributions. We used two variations of a data set: one without distribution problems (henceforth *clear* distribution), and another one in which we introduced faults in the data distribution (henceforth *confusing* distribution), as described in Section III-D.

A. Data set

We used the IMDb[1] data set, which comprises information on music, cinema, TV series, TV commercials and video games. We selected data from 2001 to 2016, resulting in 3,722 objects, and the following variables: age, genre, rating, raw profit, budget, number of Facebook likes, and the main production company. The production company is not an information from the original data base; it was obtained through automatic data mining using the movie name in the IMDb web site.

B. Questionnaire structure

The instrument of the study was an online questionnaire. First, we presented participants with an overview of the study, informing the objective, procedure, data collection, and guarantee of anonymity. We asked for their consent to use the collected answers in our research. Second, we asked the participant to answer some profile questions and report, in 7-point Likert scales: (i) their frequency in creating and analyzing charts; (ii) their familiarity with each visualization used in the study; and (iii) their knowledge about numeric data distribution concepts, linear correlation and outlier detection.

In the main part of the questionnaire, for each combination of task, chart type, and data distribution, we showed the participant one visualization and task-related (data-driven) question at a time. Visualizations were static, generated using the language R and the package ggplot2. For every question we added a checkbox "The chart does not allow me to answer", to allow us to capture the participants' assessment of the inadequacy of the chart. Figure 2 shows a fragment of a question about the chart shown in Figure 1A (or B).

The task-related questions were mandatory and had a range of possible answer formats: general text field, True/False multiple choice, and non-exclusive multiple choice (with an added option "None"). In addition to the task-related question, we included two Likert-scale statements: about the participants' confidence level in their answers, motivated by existing work [44], [45]; and about the perceived quality of the visualization to perform the task. Participants could also make additional (optional) comments in an open text field. We collected the response times for each question. Through our questionnaire, we aimed to verify whether and how data distribution might affect participants' answers for each combination of task, chart type, and distribution.

We used a total of 39 question-visualization pairs, each with 2 variations – clear and confusing data set – in a total of 78 items. We split the tasks and corresponding items (<question, chart type, distribution>) in two subsets, so that the evaluation would be less exhausting. Each participant responded the questions, in random order, of one of the subsets. After answering all questions in the first subset, the participant was asked whether he/she would like to answer an additional subset of questions, which were also randomized.

Existing work on the evaluation of visualization effectiveness aims to assess whether a visualization is capable of supporting or improving a person's answer to a certain analysis question. They usually measure success/failure rates and the time required to conduct the tasks. However, as most of these studies focus on the volume of answers (and often through crowdsourcing), their questionnaires are created to optimize the participants' time by simplifying the data set (*e.g.*, with only 5-34 data points [14]) and the question design itself (*e.g.*, multiple choice with only two possible answers [21]).

To ensure that our questionnaire would be accessible to color-blind participants, we used the color scheme proposed by Okabe and Ito [46].

C. Selection of visualization tasks, chart types, and corresponding questions

We used the relationship between tasks and visualizations proposed in [20], and included boxplot, as it has been considered a good visualization tool to analyze data distributions [47]. Table I shows the relationship between the selected charts for each type of visualization task. To keep the length of the questionnaire reasonable, we did not use georeferenced or hierarchical data in our study, nor the corresponding charts.

In a previous study, inspired by the work of Stasko and Amar [48], we asked a group of people to write questions about the IMDb online data set. We obtained a total of 76 questions, which we organized according to the tasks shown in Table I. For each task, we chose a representative question from that pool of questions, as follows:

RV: How many Action movies were released in 2015?
FE: In which year was Comedy's raw profit minimal?
MC: In which years was there greater loss than profit?

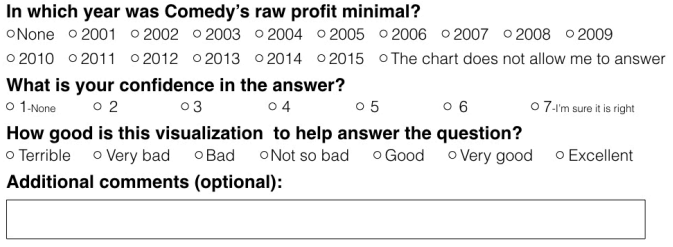

Figure 2. Question example for chart Figure 1 A (or B) [task: Find Extremum, type: Bar chart]

[1]https://www.imdb.com

Table I
RELATIONSHIP BETWEEN SELECTED TASKS AND VISUALIZATIONS (ADAPTED FROM [20])

Task types	Chart types
Return value (RV)	Bar, Line, Area, Pie, Stacked bar, Stacked area, Scatterplot, Bubble
Find extremum (FE)	Bar, Line, Area, Pie, Stacked bar, Stacked area, Scatterplot, Bubble
Make comparisons (MC)	Bar, Line, Area, Pie, Stacked bar, Stacked Area, Scatterplot, Bubble
Determining range (DR)	Bar, Line, Stacked area, Scatterplot, Bubble
Find correlation (FC)	Line, Area, Scatterplot, Bubble
Characterize distribution (CD)	Histogram, Boxplot
Find anomalies (FA)	Histogram, Boxplot, Scatterplot, Bubble

DR: In which year was the budget interval (MAX - MIN) the greatest?
FC: Is there a linear relation between Facebook Likes and IMDb score?
CD: What is the type of distribution of the number of critics for Universal Pictures?
FA: Which genre has outliers (extreme data points), if any?

As we used the same question for all chart types in each visualization task, we conducted two adjustments to ensure that the answer would not be the same across visualizations: we multiplied the analysis values by a random factor and/or switched the years randomly. This means that each visualization had its own subset of data, with the same distribution shape as the original, but with different values. To evaluate the effectiveness of the chart types given different distributions, for each original subset we created a confusing subset by inserting some disturbances in the data, as we explain next.

D. Distribution disturbances

As we wanted to evaluate whether the effectiveness of a visualization is independent of the data distribution, we added some disturbances to the original data subsets to create *confusing* data sets. We added three types of disturbances that represent real problems in data distribution, as in [49]:

- Peaks: We randomly chose a variable in the set and increased or decreased its value by 70% of the greatest value in the set, *e.g.*, in bar charts.
- Gaps: We randomly chose a value from the set and removed its n closest data points, *e.g.*, in histograms.
- Anomalies: We randomly added n points (*e.g.*, in histograms and boxplots) with values within [MIN, Q1 - 1.5*IQR] or [Q3 + 1.5*IQR, MAX], where: MIN is the smallest value, MAX is the largest value, Q1 is the value in the first quartile, Q3 is the value in the third quartile, and IQR is the interquartile range.

For each <task, chart>, we created the *confusing* distribution by applying a single disturbance, to avoid confounding the results or considerably increasing the length of the survey to deal with all possible disturbance combinations.

IV. RESULTS

The questionnaire was available for 15 days. We obtained 119 accesses, but only included the answers from the 50 participants who answered all the questions of at least one of the groups. Participants took, on average, 35 min to answer one group of questions, and 56 min to answer both groups.

A. Participant profiles

Most participants (42) were between 21 and 30 years old, 3 younger than 21, and 5 over 50. Twenty-six had Bachelor's degrees, 6 had some specialization, 16 had Master's degrees, and 4 had Doctorate degrees. Only 6 participants had formal education in the Humanities; all others came from STEM fields, mostly from IT or Engineering. Most respondents create (28 participants) and read (31) charts frequently (score 6 or 7 in a 1-7 scale). Few participants knew little to nothing (score 1 or 2 in a 1-7 scale) about numeric data distribution concepts (4), outlier detection (6), and linear correlations (6). Regarding chart types, the best known were Bar, Line, Pie, and Histogram (median M=6, interquartile range IQR=1, in a 1-7 scale), followed by Stacked bars (M=6, IQR=2); Area, Bubble, and Stacked area (M=5, IQR=2); Scatterplot (M=5, IQR=3); and Boxplot (M=4; IQR=2) (Figure 3).

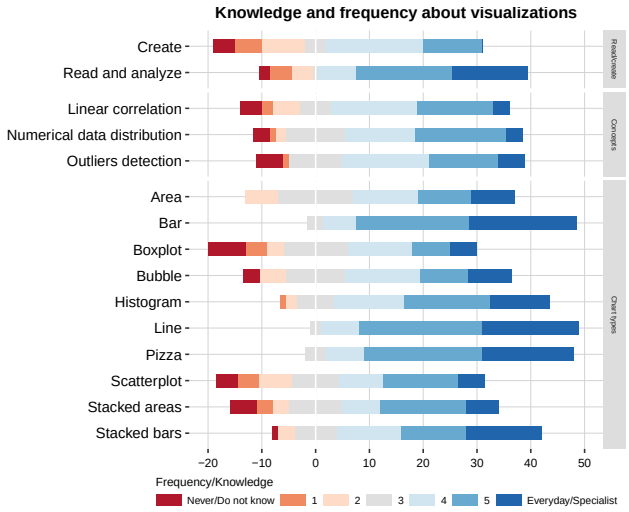

Figure 3. Participants' knowledge about visualization types and concepts, and frequency of reading and creating visualizations.

B. Chart effectiveness per task

One of our hypotheses was that the effectiveness of the chart (*i.e.*, percentage of correct answers) for each task would be related to the data distribution (clear $\times$ confusing). To evaluate this hypothesis, we conducted, for each task-chart pair, a Fisher's exact test (FET) [50], ideal for small samples or for when observed or expected values are less than 5.

Table II shows the effectiveness of each chart type for each task, for both clear and confusing distributions. It also shows the p-value resulting from the FET, and an indication of the test significance.

Table II
EFFECTIVENESS (% OF CORRECT ANSWERS)

task/chart type	clear	confusing	p-value	sig
Characterize distribution (CD)				
Histogram	**66.67%**	12.12%	1.02e-05	**
Boxplot	51.52%	3.03%	1.15e-05	**
Determine range (DR)				
Stacked area	**93.94%**	12.12%	6.80e-12	**
Bar	**93.94%**	3.03%	4.98e-15	**
Line	**90.91%**	51.52%	8.11e-04	**
Scatterplot	**81.82%**	42.42%	2.02e-03	**
Bubble	**63.64%**	15.15%	1.12e-04	**
Find anomalies (FA)				
Histogram	**84.85%**	12.12%	2.84e-09	**
Boxplot	**75.76%**	42.42%	1.16e-02	*
Scatterplot	**63.64%**	51.52%	4.55e-01	
Bubble	36.36%	6.06%	5.35e-03	**
Find correlations (FC)				
Line	**89.19%**	0.00%	1.16e-16	**
Area	**72.97%**	18.92%	5.67e-06	**
Bubble	**72.97%**	18.92%	5.67e-06	**
Scatterplot	**72.97%**	5.41%	1.51e-09	**
Find extremum (FE)				
Area	**100.00%**	**75.68%**	2.25e-03	**
Bar	**100.00%**	16.22%	6.98e-15	**
Line	**97.30%**	**67.57%**	1.38e-03	**
Stacked area	**91.89%**	29.73%	4.40e-08	**
Scatterplot	**89.19%**	5.41%	5.64e-14	**
Bubble	**75.68%**	35.14%	9.21e-04	**
Stacked bar	**72.97%**	29.73%	4.07e-04	**
Make comparisons (MC)				
Bar	**81.82%**	0.00%	9.04e-13	**
Line	**81.82%**	0.00%	9.04e-13	**
	81.82%	0.00%	9.04e-13	**
Stacked bar	**75.76%**	3.03%	5.59e-10	**
Scatterplot	**60.61%**	9.09%	1.91e-05	**
Bubble	54.55%	6.06%	2.83e-05	**
Area	30.30%	0.00%	8.77e-04	**
Stacked area	27.27%	0.00%	2.08e-03	**
Retrieve value (RV)				
Bar	**91.89%**	18.92%	1.43e-10	**
Line	**86.49%**	27.03%	3.58e-07	**
Area	**86.49%**	2.70%	2.84e-14	**
Stacked area	**83.78%**	5.41%	2.75e-12	**
Scatterplot	**75.68%**	16.22%	4.47e-07	**
Bubble	**72.97%**	0.00%	5.93e-12	**
Stacked bar	**64.86%**	13.51%	1.08e-05	**

Fisher's exact test results: ** means $p < 0.01$ and * means $p < 0.05$.

Assuming an arbitrary threshold of 60% correct answers to consider a chart as good for a certain task, we found that some chart types did not perform well, even with the clear distribution. The Boxplot performed worse than the Histogram in the *Characterize the distribution* task. This was somewhat expected, given the participants' self-reported knowledge levels (Figure 3) and given that histograms convey more information than boxplots. The Bubble chart also did not perform well for the task of *Finding anomalies (outliers)*, even in the clear distribution, and even though it had a similar structure as the Scatterplot, plus a third variable (unrelated to the task) mapped onto the size of the bubbles. We hypothesize that the inclusion of visual clutter from the different sizes of the bubbles may have caused the difference in performances between the Scatterplot and the Bubble chart, but this requires further studies. The task of *Making comparisons* also had some ineffective chart types: Bubble, Area, and Stacked Area. Analyzing the comments, we have identified that people confused the Area and Stacked area charts: "*I cannot even tell whether it is stacked or whether the vertical value starts from the horizontal axis.*" Comparing the results for the two types of distribution, we verified that there is a significant difference in effectiveness in all cases except one: Scatterplot for *Finding anomalies*.

C. Chart efficiency per task

Regarding the efficiency of each task-chart pair, the task duration was not normally distributed, so we used Mann-Whitney tests to compare the response times for correct answers across distributions. In most cases, the response time did not differ significantly. We only found significant differences in three cases, in which participants took significantly longer to provide a correct answer with the confusing (co) distribution than with the clear (cl) distribution, shown with their medians below.

- DR, Scatterplot: $M_{cl} = 29.5, M_{co} = 54, p = 4.42e^{-03}$
- FC, Area: $M_{cl} = 24, M_{co} = 46, p = 4.03e^{-02}$
- FE, Bubble: $M_{cl} = 34.5, M_{co} = 52, p = 2.02e^{-02}$
- FE, Line: $M_{cl} = 23, M_{co} = 36, p = 1.69e^{-02}$

D. User preferences

For the charts with similar effectiveness (see Table II), we analyzed, through Mann-Whitney tests, whether there was a significant difference in user preference between each pair of charts, as expressed by participants' ratings on how adequate each chart was for answering the corresponding question.

For *Determine range*, Stacked area and Bar had the same percentage of correct answers, but Bar received higher ratings than Stacked area, with $p = 0.002$. For *Find correlations or trends*, although Area and Scatterplot had the same percentage of correct answers, Area received higher ratings than Scatterplot, with $p = 0.040$. For *Make comparisons*, Bar received higher ratings than Pie, with $p = 0.015$.

E. Pairwise comparison of chart effectiveness for each task

In an attempt to rank the charts in terms of effectiveness for the same task, we compared the percentage of correct answers across pairs of charts. Table III shows the result for significant cases for clear, confusing, and all distributions.

Analyzing the *clear* distribution, for *Determine range*, Bar, Line, and Stacked area were all better than Bubble; for *Find anomalies*, Boxplot and Histogram were better than Bubble; for *Find extremum*, Area, Bar, and Line were better than Bubble or Stacked bar; for *Make comparisons*, Bar, Line, and Stacked Bar were better than Area or Stacked area; and for *Retrieve Value*, Bar was better than Bubble and Stacked bar. When we analyze the cases independent of the distribution (column *all*), we see a different picture. This means that a chart that works for *clear* distributions may not work as well for any distribution.

After these analyses, we can recommend chart types for each visualization task, as shown in table IV. The charts in the *avoid* column fared significantly worse than their counterparts.

Table III
COMPARING CHART EFFECTIVENESS

better	worse	all	clear	confusing
Determine range (DR)				
Bar	Bubble		**5.35e-03**	
Line	Bubble	**4.63e-04**	1.69e-02	
Stacked area	Bubble		**5.35e-03**	
Find anomalies (FA)				
Boxplot	Bubble		**2.92e-03**	
Histogram	Bubble		**1.58e-04**	
Find extremum (FE)				
Area	Bar	**1.02e-04**		**9.66e-07**
Area	Bubble	**2.75e-05**	**2.25e-03**	**1.06e-03**
Area	Scatter	**3.54e-07**		**3.59e-10**
Area	Stacked area	**3.50e-04**		**1.95e-04**
Area	Stacked bar	**3.38e-06**	**9.70e-04**	**1.95e-04**
Bar	Bubble		**2.25e-03**	
Bar	Stacked bar		**9.70e-04**	
Line	Bar	**2.23e-03**		**2.22e-05**
Line	Bubble	**7.40e-04**	1.38e-02	1.05e-02
Line	Scatter	**1.67e-05**		**2.16e-08**
Line	Stacked area	**6.24e-03**		**2.50e-03**
Line	Stacked bar	**1.22e-04**	**6.58e-03**	**2.50e-03**
Make comparisons (MC)				
Bar	Area		**7.24e-05**	
Bar	Stacked area		**2.64e-05**	
Line	Area		**7.24e-05**	
Line	Stacked area		**2.64e-05**	
Stacked bar	Area		**5.55e-04**	
Stacked bar	Stacked area		**2.20e-04**	
Retrieve value (RV)				
Bar	Stacked bar		**9.55e-03**	

It is important to note that here we have focused only on the data distributions. Other characteristics may influence the chart choice, as extensively discussed elsewhere (*e.g.*, [51]–[54]).

Table IV
RANKING OF CHART TYPES ACCORDING TO EFFECTIVENESS

task	consider using	avoid
Characterize distribution	Histogram	
Determine range	Stacked area, Bar, Line	Bubble
Find anomalies	Histogram, Boxplot, Scatterplot	Bubble
Find correlations	Line, Area, Bubble, Scatterplot	
Find extremum	Area, Bar, Line, Stacked area, Scatterplot	Bubble, Stacked bar
Make comparisons	Bar, Line,	Area, Stacked area, Bubble
Retrieve value	Bar, Line, Area, Stacked area, Scatterplot	Bubble, Stacked bar

F. Effectiveness vs. rating, confidence and previous knowledge

Concerning the self-reported knowledge about each chart and to the rating and confidence level of each answer, as these measures have ordinal scales, we ran Mann-Whitney tests.

Regarding the **previous knowledge about each chart**, higher prior knowledge was related to getting the answer correct in only three cases, all with the confusing distribution: Boxplot and Scatterplot for *Finding anomalies* ($p = 0.018$ and $p = 0.008$) and Stacked area for *Finding extremum* ($p = 0.042$).

Regarding the **participants' confidence level**, we analyzed the data from all distributions of three groups: participants who answered correctly (*y*), incorrectly (*n*), and who stated the chart did not answer the question (*dna*). We found a significant difference with a Kruskal-Wallis test ($p = 4.06e - 63$), so we ran post-hoc Mann-Whitney tests with Bonferroni correction, and found a significant difference in all three cases:

- $y \times n$: $M_y = 6, M_n = 5, p = 3.98e^{-34}$
- $y \times dna$: $M_y = 6, M_{dna} = 4, p = 1.12e^{-31}$
- $n \times dna$: $M_n = 5, M_{dna} = 4, p = 2.88e^{-04}$

It is worth noting that the lowest confidence level occurred when the participants believed the chart did not answer the question, even lower than when they got the answer wrong. By contrast, when analyzing the confusing distribution alone, there are only significant differences between the groups (y,n), but not between (y,dna) nor (n,dna). In other words, the underlying data distribution affected the participants' confidence level in their answers.

Regarding the **participants' rating**, we performed an analogous analysis, with similar results. When analyzing data from all distributions, we found a significant difference with a Kruskal-Wallis test ($p = 1.20e - 128$), so we ran post-hoc Mann-Whitney tests with Bonferroni correction, and found a significant difference in all three cases:

- $y \times n$: $M_y = 5, M_n = 4, p = 6.04e^{-12}$
- $y \times dna$: $M_y = 5, M_{dna} = 2, p = 3.92e^{-145}$
- $n \times dna$: $M_n = 4, M_{dna} = 2, p = 2.19e^{-94}$

As expected, when the participants believed the chart did not answer the question, they rated it as inadequate (median 2 in a 1-7 scale). Similar results were found when analyzing the clear and the confusing distributions separately, with significant differences in all three pairs of comparisons.

V. DISCUSSION

Saket et al. [14] defined five guidelines to help choose which visualization type to use, based on time on task, accuracy (effectiveness) and user preferences (rating). They were:

G1: Use bar charts for finding clusters;
G2: Use line charts for finding correlations;
G3: Use scatterplots for finding anomalies;
G4: Avoid line charts for tasks that require readers to precisely identify the value of a specific data point;
G5: Avoid using tables and pie charts for correlation tasks.

In their study, Line chart performed better than Scatterplot in all measured variables (G2). However, we did not find significant differences between Line and Scatterplot, in any distribution, regarding either accuracy or user preference. Regarding time on task, our results go in the opposite direction: Scatterplot performed significantly better than Line. This discrepancy suggests that additional studies need to be conducted to further explore these charts.

In contrast to G3, in our study Histograms and Boxplots were more effective than Scatterplots, although Scatterplots were more effective ($p = 4.19e^{05}$) and were rated higher

($p = 0.004$) than Bubble charts, regardless of the distribution. There was no significant difference on time on task between Scatterplots and the other charts for this task.

Despite G4, in our study, Line charts were highly effective for all the tasks in which they were tested (DR, FC, FE, MC, RV) with the clear distribution. Moreover, our pairwise comparison of effectiveness shows that Line is significantly better than many other charts type for *Find extremum* and *Make comparisons*, regardless the distribution. Regarding user preferences, Line received significant higher rating in several cases, for example, compared to Scatterplot ($p = 0.045$), Stacked area ($p = 0.02$), and Stacked bar ($p = 0.04$), for *Retrieve value*, regardless the distribution. Our study showed no significant difference in time on task involving Line charts.

Guidelines G1 and G5 lie outside the scope of our work, because we did not investigate the *Finding clusters* task, nor did we use Tables or Pie charts for *Finding correlation*.

VI. Threats to Validity

In this section we discuss some of the threats to the validity of our study and the actions taken to mitigate them. The data used in each chart may not have characterized the distributions as desired. To mitigate this threat, we used the same set of data with minor manipulations for the clear distribution, and with disturbances previously investigated in [49] to generate the confusing distributions. As the instrument of the study was an online questionnaire, we cannot guarantee that there was no guessing. To mitigate this threat, we measured the time on task and compared the answers within and across participants.

The chart elements and design choices may have influenced the results. To mitigate this threat, we adopted best practices from the literature when designing the charts. There was one exception: in the Bubble chart, for the confusing case, the variable to be analyzed was mapped onto the bubble size, which may have contributed to reducing the effectiveness of the chart even more.

For Stacked Bars and Stacked Areas, when the analysis variable was higher on the stack, away from the X-axis, performance decreased significantly when comparing it with mappings of the same variable next to the axis. The difference in the stack order of the target data should have been the same in both distributions, to eliminate a potentially confounding variable.

Following good chart design practices, however, led us to make a mistake: we added data labels to the Pie charts, but this prevented us from gathering data about its effectivness and efficiency as a chart, for participants needed only to read the values, without considering the graphical representation *per se*. We then decided to discard the results regarding Pie charts, so as not to distort the analysis and reach invalid conclusions.

VII. Conclusion

We conducted an empirical study to assess the effectiveness (accuracy), efficiency (time on task) and user preference (rating) to identify which types of visualization better support certain visualization tasks. We used seven different tasks, ten chart types, and two variations of a data set (clear and confusing distributions). We set out to verify whether and how data distribution affects participants' answers for each <task, chart type, distribution>. Comparing the results of the two types of distribution, we verified that there is a significant difference in effectiveness in all cases except one: Scatterplot for *Finding anomalies* which, although it had a good result with the clear distribution, the difference was not significant. For *Finding extremum*, although Area and Line charts had significantly different effectiveness across distributions, in both cases their effectiveness was deemed good ($\geq 60\%$).

With this study, we were able to identify some charts that perform better according to the task, regardless of the distribution. Our results show Area charts and Scatterplots are good to *Find correlations*, but people prefer Area charts over Scatterplots for this task. For *Make comparisons*, Bar was the most effective chart.

We also identified that the Bubble chart is not recommended for *Retrieve value* when the analysis variable is mapped onto the size of the bubble. Likewise, Stacked area is not recommended for *Make comparisons* when the analysis category is not close to the axis. Moreover, participants could not *Find extremum* using Scatterplots with non-proportional axes.

Most results pointed to a significant difference between effectiveness, confidence, and rating across distribution (clear *vs.* confusing). This calls for further comprehensive studies, as well as combining different disturbances in each pair <task, chart>, to derive more fine-grained recommendations. As an extension of this work, for each task and visualization with good performance for clear datasets, we are evaluating possible solutions for handling confusing distributions, aiming at high effectiveness, confidence, and rating.

Acknowledgment

We thank all the anonymous participants of our study. We also thank CAPES and CNPq for the financial support to this work.

References

[1] M. Card, *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[2] S. Gnanamgari, "Information presentation through default displays," Ph.D. dissertation, The Wharton School, Univ. of Pennsylvania, Philadelphia, Pa., may 1981.

[3] J. Mackinlay, "Automating the design of graphical presentations of relational information," *ACM Transactions On Graphics*, vol. 5, no. 2, pp. 110–141, 1986.

[4] P. Hanrahan, "Vizql: a language for query, analysis and visualization," in *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*. ACM, 2006, pp. 721–721.

[5] S. F. Roth and J. Mattis, "Data characterization for intelligent graphics presentation," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1990, pp. 193–200.

[6] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *The Craft of Information Visualization*. Elsevier, 2003, pp. 364–371.

[7] F. B. Viegas, M. Wattenberg, F. Van Ham, J. Kriss, and M. McKeon, "Manyeyes: a site for visualization at internet scale," *IEEE Trans. Visual Comput. Graphics*, vol. 13, no. 6, pp. 1121–1128, 2007.

[8] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer, "Voyager: Exploratory analysis via faceted browsing of visualization recommendations," *IEEE Trans. Visual Comput. Graphics*, vol. 22, no. 1, pp. 649–658, 2016.

[9] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer, "Vega-lite: A grammar of interactive graphics," *IEEE Trans. Visual Comput. Graphics*, vol. 23, no. 1, pp. 341–350, 2017.

[10] A. Key, B. Howe, D. Perry, and C. Aragon, "Vizdeck: self-organizing dashboards for visual analytics," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, 2012, pp. 681–684.

[11] M. Vartak, S. Huang, T. Siddiqui, S. Madden, and A. Parameswaran, "Towards visualization recommendation systems," *ACM SIGMOD Record*, vol. 45, no. 4, pp. 34–39, 2017.

[12] T. A. F. de Sousa and S. D. J. Barbosa, "Sistema de recomendação para apoiar a construção de gráficos com dados estatísticos," in *Proceedings of the 12th Brazilian Symposium on Human Factors in Computing Systems*. Brazilian Computer Society, 2013, pp. 168–177.

[13] B. Ondov, N. Jardine, N. Elmqvist, and S. Franconeri, "Face to face: Evaluating visual comparison," *IEEE Trans. Visual Comput. Graphics*, vol. 25, no. 1, pp. 861–871, 2019.

[14] B. Saket, A. Endert, and C. Demiralp, "Task-based effectiveness of basic visualizations," *IEEE Trans. Visual Comput. Graphics*, 2018.

[15] C. Ware, *Information visualization: perception for design*. Elsevier, 2012.

[16] J. Bertin, W. J. Berg, and H. Wainer, *Semiology of graphics: diagrams, networks, maps*. University of Wisconsin press Madison, 1983.

[17] W. S. Cleveland and R. McGill, "Graphical perception: Theory, experimentation, and application to the development of graphical methods," *Journal of the American statistical association*, vol. 79, no. 387, pp. 531–554, 1984.

[18] J. Heer and M. Bostock, "Crowdsourcing graphical perception: using mechanical turk to assess visualization design," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2010, pp. 203–212.

[19] J. Mackinlay, P. Hanrahan, and C. Stolte, "Show me: Automatic presentation for visual analysis," *IEEE Trans. Visual Comput. Graphics*, vol. 13, no. 6, pp. 1137–1144, 2007.

[20] S. Lee, S.-H. Kim, and B. C. Kwon, "Vlat: Development of a visualization literacy assessment test," *IEEE Trans. Visual Comput. Graphics*, vol. 23, no. 1, pp. 551–560, 2017.

[21] Y. Kim and J. Heer, "Assessing effects of task and data distribution on the effectiveness of visual encodings," in *Computer Graphics Forum*, vol. 37, no. 3. Wiley Online Library, 2018, pp. 157–167.

[22] V. F. de Santana, F. M. de Moraes, and B. S. R. de Souza, "Investigating treemap visualization in inverted scale," in *Proceedings of the 14th Brazilian Symposium on Human Factors in Computing Systems*. ACM, 2015, p. 34.

[23] V. F. de Santana, R. A. de Paula, and C. S. Pinhanez, "Modeling task deviations as eccentricity distribution peaks," in *Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems*. ACM, 2017, p. 36.

[24] H.-J. Schulz, T. Nocke, M. Heitzler, and H. Schumann, "A design space of visualization tasks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2366–2375, 2013.

[25] S. Wehrend and C. Lewis, "A problem-oriented classification of visualization techniques," in *Proceedings of the First IEEE Conference on Visualization: Visualization90*. IEEE, 1990, pp. 139–143.

[26] M. X. Zhou and S. K. Feiner, "Visual task characterization for automated visual discourse synthesis," in *CHI*, 1998.

[27] S. Kerpedjiev, G. Carenini, N. Green, J. Moore, and S. Roth, "Saying it in graphics: from intentions to visualizations," in *Proceedings IEEE Symposium on Information Visualization (Cat. No. 98TB100258)*. IEEE, 1998, pp. 97–101.

[28] E. H.-h. Chi and J. T. Riedl, "An operator interaction framework for visualization systems," in *Proceedings IEEE Symposium on Information Visualization (Cat. No. 98TB100258)*. IEEE, 1998, pp. 63–70.

[29] P. Pirolli and S. Card, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," in *Proceedings of international conference on intelligence analysis*, vol. 5. McLean, VA, USA, 2005, pp. 2–4.

[30] B. Lee, C. Plaisant, C. S. Parr, J.-D. Fekete, and N. Henry, "Task taxonomy for graph visualization," in *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*. ACM, 2006, pp. 1–5.

[31] M. Brehmer and T. Munzner, "A multi-level typology of abstract visualization tasks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2376–2385, 2013.

[32] H. Lam, M. Tory, and T. Munzner, "Bridging from goals to tasks with design study analysis reports," *IEEE Trans. Visual Comput. Graphics*, vol. 24, no. 1, pp. 435–445, 2018.

[33] R. Amar, J. Eagan, and J. Stasko, "Low-level components of analytic activity in information visualization," in *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* IEEE, 2005, pp. 111–117.

[34] D. A. Szafir, "Modeling color difference for visualization design," *IEEE Trans. Visual Comput. Graphics*, vol. 24, no. 1, pp. 392–401, 2018.

[35] A. Sarikaya and M. Gleicher, "Scatterplots: Tasks, data, and designs," *IEEE Trans. Visual Comput. Graphics*, vol. 24, no. 1, pp. 402–412, 2018.

[36] A. Srinivasan and J. Stasko, "Natural language interfaces for data analysis with visualization: Considering what has and could be asked," in *Proceedings of EuroVis*, vol. 17, 2017, pp. 55–59.

[37] Y. Chen, J. Yang, and W. Ribarsky, "Toward effective insight management in visual analytics systems," in *2009 IEEE Pacific Visualization Symposium*. IEEE, 2009, pp. 49–56.

[38] D. Skau, L. Harrison, and R. Kosara, "An evaluation of the impact of visual embellishments in bar charts," in *Computer Graphics Forum*, vol. 34, no. 3. Wiley Online Library, 2015, pp. 221–230.

[39] D. Albers, M. Correll, and M. Gleicher, "Task-driven evaluation of aggregation in time series visualization," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2014, pp. 551–560.

[40] J. Heer, N. Kong, and M. Agrawala, "Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2009, pp. 1303–1312.

[41] M. Siegrist, "The use or misuse of three-dimensional graphs to represent lower-dimensional data," *Behaviour & Information Technology*, vol. 15, no. 2, pp. 96–100, 1996.

[42] I. Spence and S. Lewandowsky, "Displaying proportions and percentages," *Applied Cognitive Psychology*, vol. 5, no. 1, pp. 61–77, 1991.

[43] D. Toker, C. Conati, G. Carenini, and M. Haraty, "Towards adaptive information visualization: on the influence of user characteristics," in *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 2012, pp. 274–285.

[44] M. Correll and M. Gleicher, "Error bars considered harmful: Exploring alternate encodings for mean and error," *IEEE Trans. Visual Comput. Graphics*, vol. 20, no. 12, pp. 2142–2151, 2014.

[45] J. Hullman, X. Qiao, M. Correll, A. Kale, and M. Kay, "In pursuit of error: A survey of uncertainty visualization evaluation," *IEEE Trans. Visual Comput. Graphics*, vol. 25, no. 1, pp. 903–913, 2019.

[46] M. Okabe and K. Ito, "How to make figures and presentations that are friendly to color blind people," *University of Tokyo*, 2002.

[47] H. Wickham and L. Stryjewski, "40 years of boxplots," *Am. Statistician*, 2011.

[48] R. A. Amar and J. T. Stasko, "Knowledge precepts for design and evaluation of information visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 11, no. 4, pp. 432–442, 2005.

[49] M. Correll, M. Li, G. Kindlmann, and C. Scheidegger, "Looks good to me: Visualizations as sanity checks," *IEEE Trans. Visual Comput. Graphics*, vol. 25, no. 1, pp. 830–839, 2019.

[50] R. A. Fisher, "On the interpretation of χ 2 from contingency tables, and the calculation of p," *Journal of the Royal Statistical Society*, vol. 85, no. 1, pp. 87–94, 1922.

[51] J. M. Chambers, W. S. Cleveland, B. Kleiner, and P. A. Tukey, *Graphical Methods for Data Analysis*. Pacific Grove, Calif: Chapman and Hall/Cole Publishing Company, Jan. 1983.

[52] E. R. Tufte, *The Visual Display of Quantitative Information*, 2nd ed. Cheshire, Conn: Graphics Pr, Jan. 2001.

[53] S. Few, *Now You See It: Simple Visualization Techniques for Quantitative Analysis*, 1st ed. Oakland, Calif: Analytics Press, Apr. 2009.

[54] A. Cairo, *The Truthful Art: Data, Charts, and Maps for Communication*. New Riders, Feb. 2016.