Special Issue on Data Literacy: Articles

Graphical Perception of Value Distributions: An Evaluation of Non-Expert Viewers' Data Literacy

Arkaitz Zubiaga
University of Warwick, United Kingdom
Corresponding Author.
arkaitz@zubiaga.org

Brian Mac Namee
University College Dublin, Ireland
brian.macnamee@ucd.ie

An ability to understand the outputs of data analysis is a key characteristic of data literacy and the inclusion of data visualisations is ubiquitous in the output of modern data analysis. Several aspects still remain unresolved, however, on the question of choosing data visualisations that lead viewers to an optimal interpretation of data. This is especially true when audiences have differing degrees of data literacy, and when the aim is to make sure that members of a community, who may differ on background and expertise, will make similar interpretations from data visualisations. In this paper we describe two user studies on perception from data visualisations, in which we measured the ability of participants to validate statements about the distributions of data samples visualised using different chart types. In the first user study, we find that histograms are the most suitable chart type for illustrating the distribution of values for a variable. We contrast our findings with previous research in the field, and posit three main issues identified from the study. Most notably, however, we show that viewers struggle to identify scenarios in which a chart simply does not contain enough information to validate a statement about the data that it represents. In the follow-up study, we ask viewers questions about quantification of frequencies, and identification of most frequent values from different types of histograms and density traces showing one or two distributions of values.

Zubiaga, A., Mac Namee, B. (2016). Graphical perception of value distribution: an evaluation of non-expert viewers' data literacy. *The Journal of Community Informatics*, 12(3), 138—159.

Date submitted: 2015-09-30. Date accepted: 2016-06-06.

This study reveals that viewers do better with histograms when they need to quantify the values displayed in a chart. Among the different types of histograms, interspersing the bars of two distributions in a histogram leads to the most accurate perception. Even though interspersing bars makes them thinner, the advantage of having both distributions clearly visible pays off. The findings of these user studies provide insight to assist designers in creating optimal charts that enable comparison of distributions, and emphasise the importance of using an understanding of the limits of viewers' data literacy to design charts effectively.

Introduction

Although the definition of data literacy remains somewhat fluid (Koltay, 2015), (Calzada Prado and Marzal, 2013), most definitions include an ability to interpret the outputs from data analysis. For example, Harris (Harris, 2012) defines data literacy as "competence in finding, manipulating, managing, and interpreting data, including not just numbers but also text and images"; Beauchamp (Beauchamp, 2015) defines it as "the ability to interpret, evaluate, and communicate statistical information"; and Schield (Schield, 2004) as the ability "to access, assess, manipulate, summarize, and present data". Many of the outputs from data analysis referred to in these definitions take the form of data visualisations.

In fact the importance of data visualisation is included in a number of discussions on the characteristics of data literacy (Koltay, 2015), (Wright et al., 2012), (Womack, 2014). The level of literacy that members of the general public (who are not generally trained in statistics or data analytics) bring to different types of data visualisation is not always clear. Data visualisations, however, are now ubiquitous in everyday publications such as newspapers, magazines, television programmes, and online content (Heer et al., 2010). The presence of charts on Information and Communication Technologies (ICTs) is becoming more and more important as the Web is increasingly dominated by multimedia. On social media, in particular, content is often accompanied by charts and infographics to reinforce the intended message. The choice of an appropriate chart type for a particular dataset is extremely important as it can condition subsequent interpretation by viewers. Carefully selecting the chart type that most effectively allows readers to make accurate interpretations of the data is especially important when readers have differing levels of data literacy.

In this work, we conduct two user studies to assess the effectiveness of different chart types for visualising one or more distributions of values. We conduct these two user studies using a crowdsourcing platform, which enables us to survey a large and diverse set of users who are not necessarily skilful in data analytics. In the first study, we ask viewers to validate the veracity of statements about the distributions of variables shown alongside different visualisations of these distributions. Among the five types of chart compared in this first user study, we find that histograms are not only the most complete in terms of details given, but also the chart type that leads viewers to the most accurate understanding of the underlying data. We also find, however, that viewers are not good at determining the limits of what can be understood about data from different chart types, i.e. they don't know what they don't know.

In the second, follow-up study, we move on to compare two related types of charts, histograms and density traces, to assess the capacity of viewers to accurately interpret charts. This user study is in turn split into two smaller studies. In the first study, we compare viewers' ability to interpret histograms with their ability to interpret density traces when each chart shows the distribution of a single variable. In the second user study, we examine the effectiveness of different ways of visualising the distributions of two variables together in a single histogram or in a single density trace when the aim is to compare the distributions of the two variables. We compare seven different types of charts that enable comparison of distributions: histograms with overlapped, mirrored, interspersed, stacked, or cumulative bars, and density traces that are overlapped, or mirrored. This study finds that histograms lead to more accurate interpretations in both cases (i.e. showing the distributions of one or two variables), particularly when the purpose is to quantify specific frequency values. Density traces are especially helpful, on the other hand, when we want the viewer to identify the overall tendency of values within a distribution.

The findings obtained from these user studies give us deeper understanding that enables us to define guidelines that graphical designers can use to create charts that most effectively display the distributions of variables. These guidelines are intended to satisfy the graphical perception abilities of diverse communities of users, encompassing viewers of different skill sets, and making sure that the chart selected for a visualisation is correctly interpreted by as many viewers as possible.

Related Work

Research in best conveying messages extracted from charts has focused on several aspects, including automatic generation of text summaries from charts (Demir et al., 2010), (Moraes et al., 2013), identifying the core messages of charts (Corio and Lapalme, 1999), (Demir et al., 2012), adding context to charts (Heer et al., 2009), (Hullman et al., 2013), and studying perception of information from charts (Shah and Hoeffner, 2002), (Glazer, 2011). We focus on graphical perception, as the field that studies the visual decoding of information from graphical displays. One of the best known studies on chart perception is by Cleveland and McGill (Cleveland and McGill, 1984), who define a theory to examine the elementary perceptual tasks that viewers perform when looking at charts, as well as the extent to which they lead viewers to accurate understanding.

More recently both Shah and Hoeffner (Shah and Hoeffner, 2002) and Glazer (Glazer, 2011) summarise three major factors that influence viewers' interpretations of data visualisations: (i) the visual characteristics of a chart, (ii) a viewer's knowledge about charts, and (iii) a viewer's background and expectations of the content in the chart. The authors highlight, however, that no single chart type is necessarily better overall than any other, and new tasks might require careful studies to choose a suitable chart. In general, researchers have pointed out that creating appropriate charts so that viewers perceive the intended message is harder than it might at first seem, and that detailed study of the effectiveness of different chart types for different tasks is required (Friel et al., 2001), (Shah and Hoeffner, 2002). Furthermore, the literature does not contain extensive studies of how well viewers can interpret charts showing the distribution of a variable. Here we focus on the visual characteristics of a chart, and its influence on graphical perception when comparing distributions of variables.

When it comes to displaying distributions of variables with the aim of enabling comparison between distributions, numerous types of charts have been suggested in the literature. While histograms are a well-established chart type for this task (Scott, 1979), the recent tendency has moved towards boxplots and derivatives of boxplots (McGill et al., 1978). One of the best known alternatives to the standard boxplot is the violin plot (Hintze and Nelson, 1998), which is an improved version of the boxplot that incorporates a density shape, i.e., a combination of a box plot and the density trace. The boxplot is considered to be a suitable simple chart that could be easily drawn manually (Hintze and Nelson, 1998), (Muthers and Matzarakis, 2010), but that lacks detailed information on the density of a distribution. As computational tools that facilitate chart creation emerged, however, displaying density shapes in charts has gained importance because of the additional detail provided. This has led to an increase in the use of density traces, given that they are computationally easy to create and they provide the details of a distribution that cannot be seen in boxplots. In recent years, there has been substantial discussion among researchers as to whether histograms or density traces are more suitable for displaying distributions of variables for exploratory data analysis, much of which has inclined towards the use of density traces, as histograms lack detail. For instance, Silverman (Silverman, 1986) and Izenman (Izenman, 1991) argue that histograms are a traditional way to provide a visual clue of the general shape of a distribution, but that they leave much to be desired when one needs to quantify the density of an observation in a distribution of values. Scott (Scott, 2009) adds that density traces provide the essence of conveying visual information of both the frequency and relative frequencies of observations, and thus they seem more intuitively suitable for data presentation purposes. Finally, Tukey (Tukey, 1977) resorts to histograms when he intends to display a single distribution of values, but makes use of density traces when comparing two distributions of values. In this work, we look at how these two types of charts, namely histograms and density traces, affect the graphical perception of the viewer, when the goal is to acquire basic understanding of variable distributions.

Despite the high volume of research on graphical perception, we found no work studying graphical perception of multiple variable distributions in a single chart. Our work addresses this issue by comparing the ability of viewers to compare the distributions of two variables when looking at histograms and density traces, and explore different settings for an optimal visualisation. Our work also complements a recent study by Javed et al. (Javed et al., 2010) in a similar direction, who studied alternative visualisations of multiple time series, and found that separate charts are suitable for comparison across time series with a large visual span, and shared-space charts are more efficient for smaller visual spans. Our work focuses specifically on creating single charts that put together distributions.

User Study 1: Visualising a Single Distribution

In this section we describe a user study to determine the effectiveness of different chart types for illustrating the distribution of a single variable.

Experimental Method

In the basic unit in our experimental method a chart is shown to a participant along with an associated textual statement about the distribution of the variable represented in that chart. Participants rate how well they think the statement corresponds to the data represented in the

chart. This is repeated for different combinations of three different factors: (i) the underlying distribution of the variable, (ii) the type of chart used to show the data, and (iii) the type of statement made about the data. Overall, we include four different variables, five distinct chart types, and four different types of statements.

The four artificially created variables used in the study were: (i) ages of customers of an online movie service, (ii) ages of members of a youth sports centre, (iii) salaries of a city's residents, and (iv) scores of students in an exam. Each variable exhibited a different distribution. Figure 1 shows histograms illustrating the distributions of these variables. Five commonly used chart types were selected for this study: (i) bar charts showing the average value of the distribution, (ii) bee swarms, (iii) boxplots, (iv) stacked bar charts, and (v) histograms. Figure 2 shows an example of each. All of the charts shown during the user studies were created using the R programming language[1], for which we provide details to reproduce each of the charts under study.

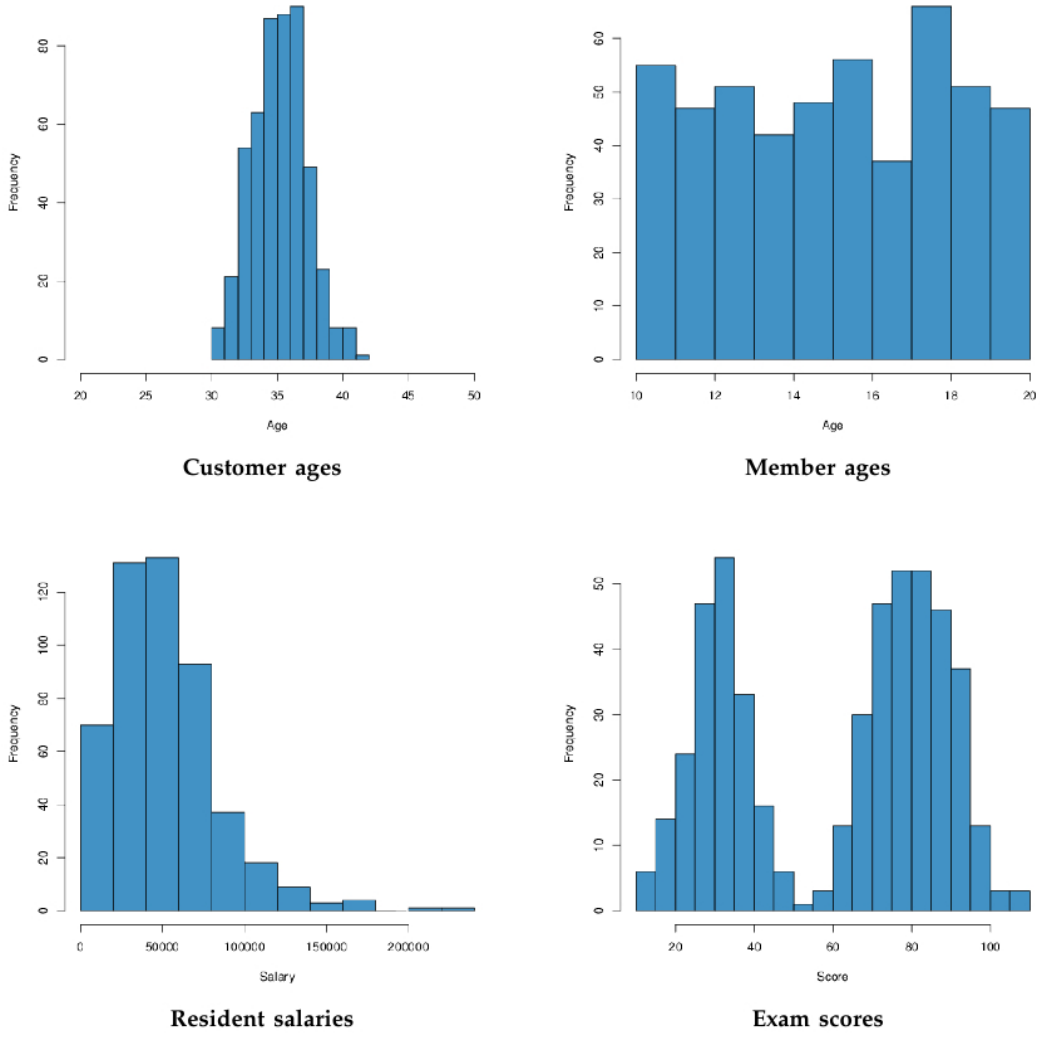


Figure 1: Histograms depicting the shapes of the data distributions under study.

[1] R: http://www.r-project.org/

The textual statements were of the following four types: (i) “the data ranges from X to Y”, (ii) “most data points fall around X”, (iii) “most data points fall under/over X”, and (iv) “data points are clustered to either side of X”. For each chart type and variable combination, two versions of each statement type were presented to participants: one that was true and one that was false. For example, for the data shown in the first histogram in Figure 1 the true statement “the data ranges from 30 to 42”, and the false statement “the data ranges from 25 to 45” were used.

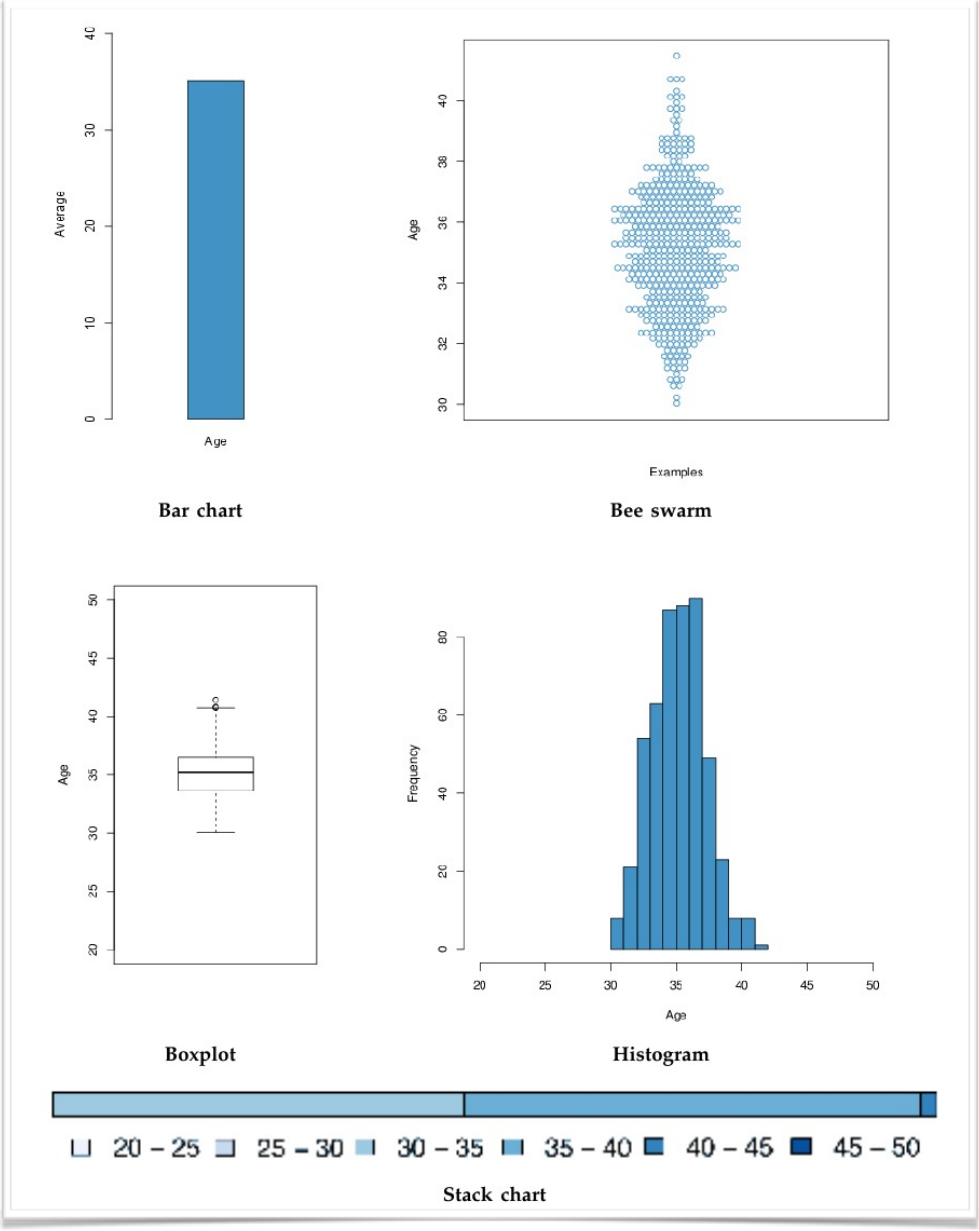


Figure 2: Examples of each chart type used for the user study.

The combinations of variables (4), chart types (5), and statements (8) amounted to a total of 160 different tasks. Participants in our experiments were shown one task at a time and had to rate the accuracy of the statement shown on a five point Likert scale: strongly agree, agree, neutral, disagree, and strongly disagree. Additionally, participants could opt for an alternative choice impossible to tell from this chart. Tasks were presented in random order to control for

learning effects. With 50 ratings collected for each of the 160 tasks, we gathered a total of 8,000 ratings.

We conducted our experiments using crowdsourcing through Mechanical Turk. The use of a crowdsourcing platform such as Mechanical Turk for this study is motivated by Heer and Bostock (Heer and Bostock, 2010), who showed that it is an effective and reliable way in which to perform graphical perception studies. To take part in the study participants did not need to have any prior expertise in data analytics as we were interested in measuring the ability of average, non-expert viewers to interpret different chart types. We did, however, restrict participation to US-based participants to control for English language capability. We also restricted participation to participants with at least a 95% HIT acceptance rate, which is Mechanical Turk's internal measure of how well participants perform tasks on the platform. A high HIT acceptance rate guarantees that participants have been deemed reliable in other experiments and filters automated bots.

Table 1: Distribution of correct answers as defined in the ground truth (T: True, F: False, I: Impossible to tell).

Statement	**T**	**F**	**I**
Data ranges from X to Y	12	12	16
Points fall around X	9	17	14
Points fall under/over X	11	13	16
Points clustered to either side of X	12	12	18
Chart Type	**T**	**F**	**I**
Bar chart (average)	0	0	32
Bee swarm	14	18	0
Boxplot	9	11	12
Stack chart	5	7	20
Histogram	14	18	0
Variable	**T**	**F**	**I**
Online movie customer ages	12	15	13
Youth sports centre ages	8	13	19
Salaries	16	16	8
Student scores	6	10	24

The ground truth for each task was manually annotated by the experiment designers, with the following distribution of responses: 42 cases were true, 54 were false, and 64 were impossible to tell. Table 1 shows the distribution of the ground truth responses, broken down by variable, chart type and statement type. The most important differences in these distributions relates to the chart types. All of the tasks showing a simple average bar chart fall into the "impossible to tell" category as the average bar chart does not provide enough evidence to assess the associated statements. With bee swarm charts and histograms it is possible, in all cases, to assess each statement. With boxplots and stacked bar charts it is possible to assess only some statements.

Results

We examine the data collected in these experiments in three ways: (1) inter-rater agreement to assess the level of agreement in the responses given by different participants; (2) accuracy to assess how well participant responses match the ground truth and (3) a confusion matrix to understand the types of errors made by participants.

Inter-rater Agreement. We measure inter-rater agreement using Krippendorff's alpha coefficient (Krippendorff, 2012). Overall, the 8,000 ratings show a fair level of inter-rater agreement of 0.392. Table 2 shows inter-rater agreement values for each chart, statement, and variable. We see two major differences here. Firstly, with regard to statement type, participants tend to agree when assessing the ranges of variables and whether variable values are above or below a given threshold; and tend to disagree when asked about values being clustered around a certain value. Secondly, with regard to chart type, participants showed a larger degree of agreement for bee swarms and histograms; and a much lower degree of agreement for the other three chart types. This is likely due to the high number of answers that are impossible to tell.

Table 2: Inter-rater agreement values by item, and overall.

Statements	
Data ranges from X to Y	0.416 (moderate)
Points fall around X	0.304 (fair)
Points fall under/over X	0.440 (moderate)
Points clustered to either side of X	0.360 (fair)
Charts	
Bar chart (average)	0.232 (fair)
Bee swarm	0.495 (moderate)
Boxplot	0.313 (fair)
Stack chart	0.211 (fair)
Histogram	0.479 (moderate)
Variables	
Online movie customer ages	0.442 (moderate)
Youth sports centre ages	0.413 (moderate)
Salaries	0.391 (fair)
Student scores	0.288 (fair)
Overall Inter-rater agreement	**0.390 (fair)**

Accuracy. To compute the accuracy values, we rely on majority voting, i.e., the rating that has been chosen by most participants. This allows us to choose a single rating from the 50 provided for each task. For the purposes of computing accuracy, we collapse ratings of agree and strongly agree to true, and ratings of disagree and strongly disagree to false[3]. The final accuracy values reported here refer to the number of cases in which the majority vote of

2 We report the strength of agreement using the benchmarks suggested by Landis and Koch (Landis and Koch, 1977) for interpreting kappa.

3 In fact, participants seemed reluctant to choose strong judgements, choosing agree and disagree much more than strongly agree and strongly disagree.

participants coincides with the ground truth. An overall accuracy value, and values broken down by variable, chart type and statement, are shown in Table 3.

Table 3: Accuracy values by item, and overall.

Statements	
Data ranges from X to Y	0.900
Points fall around X	0.550
Points fall under/over X	0.750
Points clustered to either side of X	0.525
Charts	
Bar chart (average)	0.531
Bee swarm	0.906
Boxplot	0.563
Stack chart	0.438
Histogram	0.969
Variables	
Online movie customer ages	0.700
Youth sports centre ages	0.700
Salaries	0.750
Student scores	0.575
Overall accuracy	**0.681**

On statements of the types “data ranges from X to Y” and “points fall under/over X”, participants were substantially more accurate (90% and 75%, respectively) than for the other two types of statements (55% and 52.5%). More specifically we found that participants struggled with bar and stack charts when assessing “points fall around X” statements, and with stack charts when assessing “points clustered to either side of X” statements.

Regarding the chart type, the most accurate answers were those for bee swarms and histograms (both above 90%). This is slightly surprising as these are relatively complex chart types for non-expert viewers. Even though both bee swarm charts and histograms potentially allow viewers to determine the veracity of all of the statements, and provide similar information, viewers seem to find it slightly easier to comprehend values from a histogram.

Finally, participants struggled slightly to answer questions about the student scores data. This data has a bimodal distribution that could be more difficult for viewers to parse.

Confusion Matrix. Table 4 shows a confusion matrix for all tasks (note that Imp. refers to responses of impossible to tell, that Neutral did not occur in the ground truth, and that cells marking correct responses are highlighted in bold). The precision for each category is also included. Most notably here, we observe that when the correct response was impossible to tell, participants mostly deemed statements false (45.9% of the time), or even true (24.9% of the time), and only identified 23.8% of the cases correctly. When the correct response was either true or false, participants again rarely chose impossible to tell as the answer. Taken altogether we believe that these results indicate that, although participants do well when

assessing true cases (accuracy 75.8%) and false cases (72.1%), they have trouble when facing charts that do not enable them to determine the veracity of a statement and do not recognise this shortcoming. We were surprised that participants did not use the neutral choice in these cases (the neutral response was only used in 6% of cases).

Table 4: Confusion matrix for all the tasks combined (in %).

		Responses			
		Imp.	False	Neutral	True
Ground Truth	Imp.	**23.8**	45.9	5.4	24.9
	False	6.9	**72.1**	6.4	14.6
	True	4.7	14.0	5.5	**75.8**
Precision		67.2	54.6	-	65.7

Overall, the main finding from this study was that viewers found histograms were the easiest to interpret of the five chart types studied. Using this finding, we designed a follow up study to determine the ability of viewers to compare the distributions of two variables when visualised using either histograms or density traces.

User Study 2: Expanding Charts to Visualise Multiple Distributions

We set out to conduct user studies to measure the ability of viewers to interpret differences between two variable distributions for the purpose of exploratory data analysis (Tukey, 1977). We split this exploration into two user studies. The first study investigates whether viewers find histograms or density traces easier to interpret when viewing the distribution of a single variable. Once this is established, the second study investigates which chart type is most effective for comparing the distributions of two different variables.

We also conducted these user studies through the Amazon Mechanical Turk crowdsourcing platform. Given that we wanted to conduct the user study with viewers that were not necessarily skilful in data analytics, and that we were looking at relatively simple perception tasks involving the quantification of values from charts, the use of a crowdsourcing platform presented a suitable environment for our purposes.

During both of these user studies, we set up the tasks on Mechanical Turk without restrictions on the expertise of participants in terms of their ability to decode charts. In order to make sure that participants were reliable, we again restricted participation to participants with a HIT acceptance rate of at least 95%, and also a number of completed tasks of at least 100. These settings have been found to be suitable to prevent participants who cheat (Heer and Bostock, 2010).

Our basic experimental unit consisted of showing a chart to a participant and asking them a series of questions about the chart in order to assess the accuracy with which they could interpret it. For each experimental unit we first showed the participant an entry page that displayed the chart in question, along with instructions explaining that they would be asked a series of questions about the distribution of the variable displayed in the chart. Figure 3 shows an example of this entry page. Once a participant clicked on the Start button, they were shown a question displayed next to the chart. Participants had to provide an answer to the question

and then click 'Next' to proceed to the next question. This process repeated until the participant has answered all of the questions associated with the chart. Each experimental unit was completed by 50 different participants. For each participant, we collected the answer they provided to each question, as well as the response time measured from the moment they first saw the question to the moment they clicked 'Next' having answered the question.

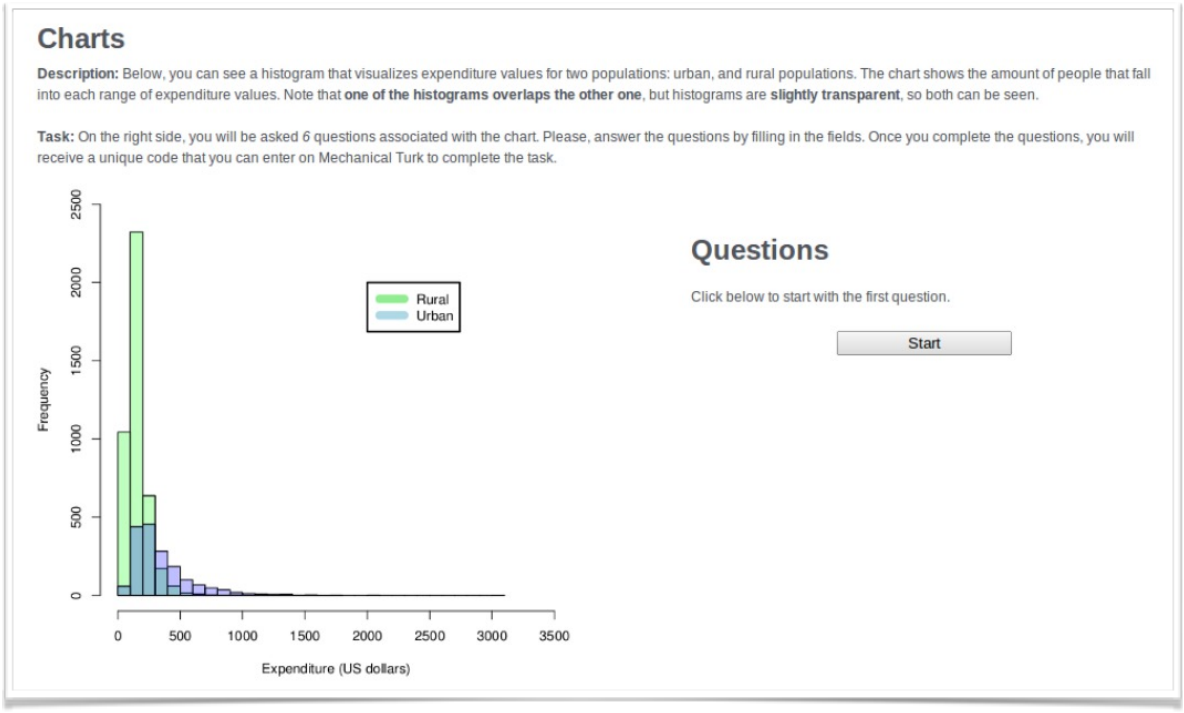


Figure 3: Entry page of the user study shown to participants.

In this paper, we report three values in the results section: (1) response time in seconds, (2) accuracy of respondents (calculated as the percentage of participants that provided an answer within a 10% error rate of the ground truth[4], and (3) the error rate for participants that were accurate according to point 2. Note that the error rate is computed as the relative difference between the correct answer and participant's answer, i.e., the deviation from the correct answer, so we can measure how close the responses were for those that were exactly or nearly accurate. We differentiate between accuracy and error rate given that our purpose is to measure both the ability of viewers to find the answer in the chart –for example, instead of mixing up axes and giving an answer for the wrong axis–, and the ability to provide a precise answer for those who found it.

The charts shown during both user studies display data from the Vietnamese Living Standard Study (VLSS), which is available online[5], and has been used for instance by Tukey (Tukey, 1977) for exploratory data analysis. The data from the VLSS contains household per capita

[4] We used 10% as a reasonable percentage to consider that participants were able to identify where to get the answer from in the chart, and their response was close enough to the correct answer. Likewise, this enables us to compare responses for histograms and density traces, which display different scales of values.

[5] VLSS Data: http://www.tc.umn.edu/~zief0002/Comparing-Groups/Data/VLSSperCapita.csv

expenditures for 5,999 Vietnamese households, divided into rural and urban areas. This enabled us to separate expenditure values into these populations, and thus to show two distributions of values that viewers had to compare. All of the charts shown during the user studies were created using the R programming language, for which we provide details to reproduce each of the charts under study.

User Study 2.1: Comparing Histograms and Density Traces for Visualising a Single Distribution

In the first of our user studies, we compare the ability of participants to interpret visualisations of the distribution of a single variable using histograms and density traces. This section describes the experimental method used for this study and the results of the study.

Experimental Method

This study used the experimental method described at the beginning of Section 4. Viewers were asked to answer questions about two different types of charts:

Histograms

Histograms (cf. (Scott, 1979), (Guha et al., 2001)) are a very commonly used way to graphically represent the distribution or a variable. A histogram shows tabulated frequency values that give the gist of how data is distributed. Given that they present tabulated frequency values, the width of each of the bins in the plot must be predefined. Among the numerous methods to define the bin width (Wand, 1997), we relied on Sturges' rule (Sturges, 1926) to create the histograms for our study. As a result, expenditure values ranging from 0 to 1800 for the urban population, and from 0 to 3100 for the rural population, were split into bin widths of 100. We used R's hist() function to plot histograms. Previously, it has been found that the orientation of the bars in a histogram has an effect on the viewers' perception. Fischer et al. (Fischer et al., 2005) concluded that vertical bars allow viewers to react quicker and make decisions faster than horizontal bars. Therefore, in this work we focus on histograms displaying vertical bars. Figures 4(b) and 4(d) show histograms displaying the distributions for expenditure values obtained from the VLSS data, one for the rural population and one for the urban population.

One advantage of histograms is that bins facilitate quantification of the frequency for each range of values. As has been pointed in the literature, however, the lack of a detailed visualisation of more points of the distribution can make accurate perception by viewers difficult.

Density traces:

Different from histograms, density traces do not use tabulated data, and instead the distribution is visualised as a continuous, single line that depicts how frequencies change across the range of possible values. To draw a density trace a kernel function is required to extract frequency values and draw the line. In this case, we rely on the commonly used kernel method introduced by Epanechnikov (Epanechnikov, 1969). We used R's density() function to plot density traces. Figures 4(a) and 4(c) show density traces displaying distributions for

expenditure values obtained from the VLSS data, one for the rural population and one for the urban population.

Density traces present the advantage over histograms that a more detailed representation of the whole distribution is shown which, it has been suggested, makes them a more suitable chart for visualising both the frequency and relative frequencies of observations. One could also expect, however, that quantification of frequencies for specific points might be difficult for viewers from a curved density trace.

With a focus on identifying the overall tendency of each distribution of values, and quantifying frequencies depicted in the charts, we asked participants to provide the following values based on interpreting the displayed charts:

- Minimum expenditure value in the distribution.
- Maximum expenditure value in the distribution.
- Most frequent value (MFV) in the distribution.
- Frequency value for an expenditure of $200.
- Frequency value for an expenditure of $500.
- Frequency value for an expenditure of $1,000.

Results

Tables 5, 6, and 7 show the average results for accuracy, error rate, and response time for histograms and density traces from this study. To compute the average response times, we remove response times above the 95th percentile, and those below the 5th percentile, to control for outliers. Finally, we also show the average accuracy values, error rates, and response times for a chart combining all the questions, which helps us assess the overall performance with each type of chart.

Note that the averages for error rates are not necessarily the arithmetic mean of the error rates for all the questions with that chart, since different number of participants might be accurate and thus be considered for computing the error rates; therefore, it represents a weighted mean for all accurate responses to each question associated with a chart.

The accuracies achieved depended on the type of question being asked. Viewers were able to more accurately use density traces when responding to questions about minimum and maximum values, as well as most frequent values. However, when responding to the other questions about frequency values viewers were more accurate when interpreting histograms. The only exception is the frequency for x = $200, where viewers were slightly more accurate, but still relatively similar, when looking at density traces. This exception in x = $200 happens to be a point with no tick mark in the X-axis, which might have made it more difficult for viewers to identify. As we hypothesised above, it appears that the fact that density traces are curved lines complicates quantification of frequency values for specific points, but facilitates identification of trends and therefore finding points such as the most frequent value.

Table 5: Accuracy values for User Study #1

	Histograms	Density Traces
Min	0.92	0.94
Max	0.38	0.54
MFV	0.78	0.92
Freq ($200)	0.69	0.43
Freq ($500)	0.78	0.56
Freq ($1000)	0.94	0.83
Average	0.748	0.703

Table 6: Error rates for User Study #1

	Histograms	Density Traces
Min	0.002	0.001
Max	0.031	0.031
MFV	0.021	0.044
Freq ($200)	0.064	0.029
Freq ($500)	0.016	0.039
Freq ($1000)	0.011	0.014
Average	0.021	0.024

Table 7: Response times (in seconds) for User Study #1

	Histograms	Density Traces
Min	29.6	32.6
Max	27.8	33.6
MFV	31.2	29.5
Freq ($200)	45.9	44.2
Freq ($500)	49.4	46.8
Freq ($1000)	48.3	46.5
Average	39.7	38.4

Overall, putting together the results for all types of questions, viewers were on average more accurate when making interpretations from histograms than from density traces. This accuracy gain with histograms is also reflected in error rates. The error rates for histograms are also slightly lower (with the only exception of the frequency for $200), which suggests histograms as a more suitable chart than density traces for viewers to make accurate interpretations.

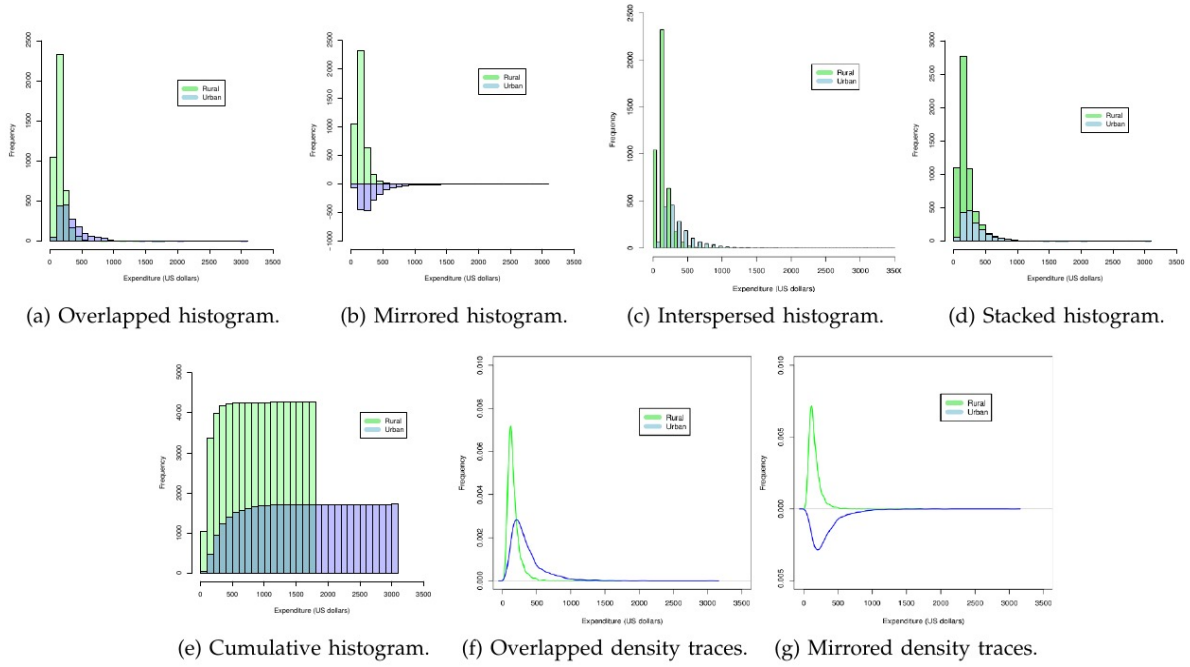
Overall, there does not seem to be a clear difference in terms of response times between histograms and density traces. On average, viewers spent only 3.4% more time in answering to questions associated with histograms, which is reflected in a 6.4% improvement in terms of accuracy.

User Study 2.2: Comparison of Two Distributions

Having seen that histograms convey more accurate interpretations than density traces when it comes to a single distribution, in a follow-up user study we looked at performance of viewers when comparing two distributions. Viewers do better in quantifying interpretations from a single histogram, but how should two histograms or two density traces be put together in a single chart to optimise perception? Since two plots can be arranged in different ways in a single chart, we study the effect of these arrangements on the perception of viewers.

Experimental Methods

This study also used the experimental method described at the beginning of Section 4. In this study viewers were asked to answer questions about six different types of charts:



(a) Overlapped histogram. (b) Mirrored histogram. (c) Interspersed histogram. (d) Stacked histogram.

(e) Cumulative histogram. (f) Overlapped density traces. (g) Mirrored density traces.

Figure 5: Histograms with different settings shown to participants of the User Study #2.

Overlapped histograms:

In an overlapped histogram, one of the histograms is superimposed on top of the other, with both lying in the X-axis (as in the first user study, based on the literature (Fischer et al., 2005), we assume that vertical bars are more suitable for viewers than horizontal bars). In order for both histograms to be seen, they are made slightly transparent. We do this by using R's hist() function including an alpha = 14 parameter in the set of colours being used. Again, we established the bin width for histograms by following Sturges' rule (Sturges, 1926) (this approach was used for all histograms in this user study). Figure 5 (a) shows the resulting overlapped histogram for the VLSS data.

We expect that overlapped histograms will enable comparison of frequencies for both distributions, but the fact that one of the distributions slightly complicates visualising the other distribution might complicate their differentiation.

Overlapped density traces:

Density traces can also be superimposed on top of each other, as shown in figure 5 (f). We used R's plot() and lines() functions combined with the density() function, using Epanechnikov's kernel (Epanechnikov, 1969), to plot these charts (this approach was used for all density traces in this user study). Having two density traces sharing the same space could aid comparison and, different from overlapping histograms, overlapping density traces do not hinder the visualisation of the lines typically occlude one another.

Mirrored histograms:

In order to avoid overlapping histograms, in a mirrored histogram one of the histograms is mirrored downwards from the X-axis. In spite of pointing downwards, the length of the bars in the bottom histogram also represent positive values. To draw mirrored histograms we used R's hist() function after inverting the values for one of the histograms. Figure 5 (b) shows the mirrored histogram for the VLSS data. We expect that mirrored histograms will facilitate clear visualisation of both distributions without any overlap, but that the quantification of bars pointing downwards from the X-axis could be more challenging for viewers.

Mirrored density traces:

Similarly, density traces can be mirrored so that one of them is drawn downwards from the X-axis. Figure 5 (g) shows mirrored density traces for the VLSS data. To draw this we used R's plot() and lines() functions after inverting the values for one of the density traces.

Similar to the advantage offered by mirrored histograms, we expect that mirrored density traces might facilitate visualisation of both lines separately avoiding possible confusion, but may make it more difficult for participants to perform comparisons between both distributions, as they do not share the same space.

Interspersed histograms:

Bars for two variables are interspersed in a single histogram, so that for each range of values two bars are shown next to each other, one for the frequency of each variable in that range. Figure 5 (c) shows the resulting interspersed histogram of the VLSS data that we showed to the participants in the study. We created this chart using the multhist() function from the 'plotrix' package in R.

We expect that interspersed histograms will facilitate visualisation of both distributions, as they do not occlude each other. This, however, is at the cost of halving the horizontal space physically available for the width of the bars, which might have a negative effect in on visual perception of viewers.

Stacked histograms:

In a stacked histogram the bars for one distribution lie on top of the bars of the other distribution. This means that the frequency values for one of the distributions do not count from the X-axis, but from an upper point on top of the bar for the other distribution. We created stacked histograms using R's histStack() function. Figure 5(d) shows the resulting stacked histogram for the VLSS data that we showed to the participants of the study.

We expect that when using stacked histograms it will be easier for viewers to differentiate histograms from each other than when overlapped histograms are used. It may however be more challenging for viewers to quantify the height of bars, as significant cognitive effort is required (viewers need to subtract the height of one bar from the other).

Cumulative histograms:

Each bar in a cumulative histogram represents the cumulative frequency for all smaller values, instead of representing just the value for that specific range. For instance, the third bar from the left for a distribution represents the aggregation of the frequencies for the first, second, and third bars. Consequently, the increase of a bar with respect to the previous bar actually represents the frequency of that specific range. Figure 5 (e) shows the cumulative histogram for the VLSS data that we showed to the participants of the study. We created this chart using R's hist() function, which received the outcome of applying the cumsum() function to the histogram's data values.

We expect that cumulative histograms will facilitate differentiation between the two distributions, but the fact that frequencies are summed will complicate quantification of specific frequencies.

Each time a visualisation was shown to participants we requested the following values as the input from participants in the user study:

- Most frequent value (MFV) for each distribution. The most frequent value is different for the rural and urban populations, and thus viewers need to precisely identify each population's most frequent value.
- Frequency values for specific data points in both distributions. We asked for the frequency for expenditure values of $200 and $500, for both populations. The main difference between these two cases is that $500 has a tick mark in the X-axis, while $200 does not. This might make a difference in the interpretation from viewers, making it potentially more difficult to position a value when there is no tick mark.

Results

Tables 8, 9, and 10 show the average results for accuracy, error rates, and response times for the charts under study.

Table 8: Accuracy values for User Study #2

	Histograms					Density Traces	
	Overlapped	Mirrored	Interspersed	Stacked	Cumulative	Overlapped	Mirrored
MFV	0.81	0.75	0.77	0.73	0.16	0.82	0.84
Freq ($200)	0.71	0.61	0.67	0.76	0.37	0.55	0.37
Freq ($500)	0.92	0.97	0.96	0.97	0.20	0.84	0.59
Average	0.77	0.72	0.85	0.78	0.21	0.70	0.58

Table 9: Error rates for User Study #2

	Histograms					Density Traces	
	Overlapped	Mirrored	Interspersed	Stacked	Cumulative	Overlapped	Mirrored
MFV	0.024	0.035	0.038	0.025	0.037	0.028	0.032
Freq ($200)	0.014	0.020	0.022	0.041	0.025	0.036	0.036
Freq ($500)	0.022	0.016	0.019	0.022	0.030	0.026	0.024
Average	0.020	0.022	0.027	0.032	0.029	0.027	0.029

Table 10: Response times (in seconds) for User Study #2

	Histograms					Density Traces	
	Overlapped	Mirrored	Interspersed	Stacked	Cumulative	Overlapped	Mirrored
MFV	24.2	25.8	25.3	25.2	23.8	23.8	27.4
Freq ($200)	20.7	23.6	23.8	23.1	18.7	21.5	23.2
Freq ($500)	19.8	22.2	17.9	20.7	22.2	20.3	21.7
Average	23.2	25.4	23.2	24.1	21.9	22.7	26.3

As expected, viewers were not able to accurately interpret cumulative histograms. The fact that the frequency for each range of values has to be calculated by subtracting the frequency for the previous range confused viewers, misleading their perception. Accuracy values of around 20% were achieved in most cases, either when looking for most frequent values, or when quantifying frequency values. Viewers were clearly more accurate with the rest of the charts, achieving average accuracies higher than 70%.

Overall viewers managed clearly better interpretations from interspersed histograms, achieving an accuracy of 85%. Displaying thinner bars gives the advantage of making both bars clearly visible without any overlap, and easily quantifiable without the need to stack bars. Still, viewers did quite well with overlapped and stacked histograms, achieving 77% and 78% accuracy rates, respectively. These two types of histograms led to better perceptions from viewers than mirrored histograms, where both distributions are visible with no overlaps. The fact that one of the distributions is mirrored downwards seems to have damaged quantification of frequency values for viewers.

If we look at the accuracy of responses by type of question, there is a noticeable lower performance when providing values for frequencies of $200 than for frequencies of $500. Again, the fact that $200 does not have a tick mark in the X-axis appears to be misleading viewers. Adding more tick marks in the X-axis as long as space allows should help boost

performance when quantifying values that are on or close to those tick marks. It is certainly key to think of the specific points in which tick marks have to be added in order to guarantee that the intended message is correctly conveyed.

With the density traces we see a similar trend as when viewers looked at a single distribution, i.e., viewers were highly accurate when identifying most frequent values (slightly more accurate even than the best of the histograms), but the performance when quantifying specific frequency values is poorer, which also drops the overall performance.

Density traces are therefore a suitable visualisation when the intention is to emphasise the central tendency of a distribution. However, histograms are more suitable when we want viewers to interpret more specific values shown in the distribution.

Looking at the response times, it can be seen that viewers needed more time to respond to questions about most frequent values, than for questions about specific frequency values. This reinforces our conclusions from the first user study that viewers seem to feel more comfortable with histograms when quantifying frequency values, but are not as comfortable when looking at the tendency of values to identify the most frequent value.

Discussion

In this work, we have conducted two user studies to assess viewers' data literacy when interpreting a distribution of values displayed in different types of charts. In the first study, we have studied the suitability of five different types of charts to visualise a single distribution of values. In a follow-up study, we have delved into different types of histograms and density traces to assess viewers' literacy not only with a single distribution of values, but also when putting two together with the aim of comparing them with each other. We have used a crowdsourcing platform to conduct these studies, without restricting users by their level of expertise, and therefore allowing participation from users with differing levels of data literacy.

In the first user study, we have seen that histograms allow the most accurate interpretations— viewers achieved 97% accuracy from histograms, compared to 91% with bee swarms, and lower than 60% for the other charts— and are an appropriate choice of chart type when visualising the distribution of a variable for an average, non-expert audience. This reinforces previous findings from Meyer et al. (Meyer et al., 1997) and Zacks and Tversky (Zacks and Tversky, 1999) concluding that bar charts are a suitable visualisation medium to support reading exact values, identification of maxima, and describing contrasts in data.

More interestingly, this study highlighted a shortcoming in the ability of average, non-expert viewers to recognise the limitations of different chart types—viewers don't know what they don't know. This is a significant issue as it means that there is a strong possibility that viewers are likely to make incorrect inferences from charts, or that they can be very easily misled using charts. This finding reinforces the need to carefully design charts for different tasks (Shah and Hoeffner, 2002), (Glazer, 2011) and highlights a shortcoming in the data literacy of non-experts.

Another interesting point arising from the apparent effectiveness of histograms compared to bee swarms is that it reinforces the finding by Fischer et al. (Fischer et al., 2005) that viewers

find it easier to interpret vertical bars (present in histograms) than horizontal bars (present in bee swarms). We also believe that there might be a difference between centring the data points in a bee swarm around a virtual vertical axis in the middle of the chart, and placing the data points upwards starting from the X-axis in a histogram. The gap between two bars lying on the same axis can be easily quantified visually, while the gap between two bars centred on an axis is halved on both sides of the bar making it more difficult to quantify. The alignment of the bars with respect to the axis might affect perception—this warrants further study.

In the second user study, our results suggest that histograms are overall more suitable than density traces to display distributions of values to viewers with different levels of expertise and not necessarily trained in data analytics, especially when the main purpose is quantification of specific frequency values. Density traces have shown instead to be more suitable to emphasise the tendency of values underlying a distribution. In a follow-up user study, we have identified that interspersing bars of the two distributions plotted in a histogram leads to optimal perception when comparing distributions. Other alternatives such as overlapping, stacking, and mirroring bars in histograms led to much less accurate perceptions, while cumulative histograms showed to be by far the worst option. The findings of these user studies provide insight towards defining guidelines to assist graphical designers in optimal creation of charts that enable comparison of distributions. The fact that our user studies have been conducted with non-expert users whose level of expertise has not been restricted makes our guidelines suitable to be applied to communities of users with different degrees of data literacy.

Future work includes deepening the comparison of value distributions, by looking into more challenging cases where three or more distributions need to be compared, given that histograms with increasing numbers of distributions might require different approaches. Another aspect that has not been dealt with in this work, and would be a sensible objective to pursue would be to break down the user study into different demographic groups to better understand how perception would affect people of different ages, cultures, etc.

Acknowledgments

This work was supported by the Enterprise Ireland and IDA Ireland Technology Centres programme at CeADAR, the Centre for Applied Data Analytics Research.

References

Beauchamp, A. (2015). What is data literacy? *Databrarians. February, 12.*

Calzada Prado, J. and Marzal, M. A. (2013). Incorporating data literacy into information literacy programs: Core competencies and contents. *Libri: International Journal of Libraries & Information Services*, 63(2):123 – 134.

Cleveland, W. S. and McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554.

Corio, M. and Lapalme, G. (1999). Generation of texts for information graphics. In *Proceedings of EWNLG'99*, 49–58.

Demir, S., Carberry, S., and McCoy, K. F. (2012). Summarizing information graphics textually. *Computational Linguistics*, 38(3):527–574.

Demir, S., Oliver, D., Schwartz, E., Elzer, S., Carberry, S., and McCoy, K. F. (2010). Interactive sight into information graphics. In *Proceedings of W4A*,16. ACM.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, 14(1):153–158.

Fischer, M. H., Dewulf, N., and Hill, R. L. (2005). Designing bar graphs: Orientation matters. *Applied Cognitive Psychology*, 19(7):953–962.

Friel, S. N., Curcio, F. R., and Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Math. Education*, 124–158.

Glazer, N. (2011). Challenges with graph interpretation: A review of the literature. *Studies in Science Education*, 47(2):183–210.

Guha, S., Koudas, N., and Shim, K. (2001). Data-streams and histograms. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, 471–475. ACM.

Harris, J. (2012). Data is useless without the skills to analyze it. *Harvard Business Review, 13*.

Heer, J. and Bostock, M. (2010). Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 203–212. ACM.

Heer, J., Bostock, M., and Ogievetsky, V. (2010). A tour through the visualization zoo. *Communications of the ACM*, 53(6):59–67.

Heer, J., Vi´ egas, F. B., and Wattenberg, M. (2009). Voyagers and voyeurs: Supporting asynchronous collaborative visualization. *Communications of the ACM*, 52(1):87–97.

Hintze, J. L. and Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician*, 52(2):181–184.

Hullman, J., Diakopoulos, N., and Adar, E. (2013). Contextifier: automatic generation of annotated stock visualizations. In *Proceedings of CHI*, 2707–2716. ACM.

Izenman, A. J. (1991). Review papers: Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, 86(413):205–224.

Javed, W., McDonnel, B., and Elmqvist, N. (2010). Graphical perception of multiple time series. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):927–934.

Koltay, T. (2015). Data literacy: in search of a name and identity. *Journal of Documentation*, 71(2): 401–415.

Krippendorff, K. (2012). Content analysis: An introduction to its methodology. *Sage*.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

McGill, R., Tukey, J. W., and Larsen, W. A. (1978). Variations of box plots. *The American Statistician*, 32(1):12–16.

Meyer, J., Shinar, D., and Leiser, D. (1997). Multiple factors that determine performance with tables and graphs. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 39(2):268–286.

Moraes, P. S., Carberry, S., and McCoy, K. (2013). Providing access to the high-level content of line graphs from online popular media. In *Proceedings of W4A*, 1–10. ACM.

Muthers, S. and Matzarakis, A. (2010). Use of beanplots in applied climatology a comparison with boxplots. *Meteorologische Zeitschrift*, 19(6):641–644.

Schield, M. (2004). Information literacy, statistical literacy and data literacy. *IASSIST Quarterly*, 28(2/3):6–11.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, 66(3):605–610.

Scott, D. W. (2009). Multivariate density estimation: theory, practice, and visualization, volume 383. *Wiley. com*.

Shah, P. and Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review*, 14(1):47–69.

Silverman, B. W. (1986). Density estimation for statistics and data analysis, volume 26. *CRC press*.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association*, 21(153):65–66.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, Ma, 231.

Wand, M. (1997). Data-based choice of histogram bin width. *The American Statistician*, 51(1):59–64.

Womack, R. (2014). *Data Visualization and Information Literacy*, volume 38.

Wright, S., Fosmire, M., Jeffryes, J., Stowell Bracke, M., and Westra, B. (2012). A multi-institutional project to develop discipline-specific data literacy instruction for graduate students. *Libraries Faculty and Staff Presentations*, Paper 10.

Zacks, J. and Tversky, B. (1999). Bars and lines: A study of graphic communication. *Memory & Cognition*, 27(6):1073–1079.