*Journal of Universal Computer Science, vol. 27, no. 10 (2021), 1046-1068*
*submitted: 30/3/2021, accepted: 29/9/2021, appeared: 28/10/2021* 

# Data-driven Storytelling to Support Decision Making in Crisis Settings: A Case Study

**Andrea Lezcano Airaldi**
(National University of NorthEast, Corrientes, Argentina,
alezcano@exa.unne.edu.ar)

**Jorge Andrés Diaz-Pace**
(UNICEN University, Tandil, Argentina,
adiaz@exa.unicen.edu.ar)

**Emanuel Irrazábal**
(National University of NorthEast, Corrientes, Argentina,
https://orcid.org/0000-0003-2096-5638, eirrazabal@exa.unne.edu.ar)

**Abstract:** Data-driven storytelling helps to communicate facts, easing comprehension and decision making, particularly in crisis settings such as the current COVID-19 pandemic. Several studies have reported on general practices and guidelines to follow in order to create effective narrative visualizations. However, research regarding the benefits of implementing those practices and guidelines in software development is limited. In this article, we present a case study that explores the benefits of including data visualization best practices in the development of a software system for the current health crisis. We performed a quantitative and qualitative analysis of sixteen graphs required by the system to monitor patients' isolation and circulation permits in quarantine due to the COVID-19 pandemic. The results showed that the use of storytelling techniques in data visualization contributed to an improved decision-making process in terms of increasing information comprehension and memorability by the system stakeholders.

**Keywords:** best practices, COVID-19, data storytelling, information visualization, empirical study
**Categories:** H.0, H.4.0, H.4.3, H.5, H.5.0
**DOI:** 10.3897/jucs.66714

## 1 Introduction

Storytelling is currently a very effective way to convey information, as its primary goal is to engage with the audience and stimulate its attention through emotions [Nussbaumer Knaflic, 2015]. When applied to data visualization, storytelling techniques help to communicate facts and facilitate their understanding [Kosara and Mackinlay, 2013]. The implementation of data storytelling best practices in the development of a software product can significantly impact the decision-making process [Schreyögg, 2006]. However, in cases where the project requirements are volatile and have strict deadlines, these practices are often set aside in favor of dealing with the minimum required functionalities [Cohen et al., 1996].

Recently, a case study by the Canadian health system stated that the COVID19 pandemic crisis has increased the speed of changes in application development [Krausz et al., 2020], in detriment of visual content quality. According to the author, there is a shortage of live visualizations and reports of data collected safely and in real-time when considering the different areas of public health and security for decision making. This kind of problems can even lead to the rejection of the development by its stakeholders. For instance, in the Netherlands, the first developments lasted weeks but were rejected due to issues in protecting personal data [Janssen and van der Voort, 2020]. In general, user interface issues seem to be very common in COVID-19 software projects [Rahman and Farhana, 2020].

Making informed decisions to mitigate the impact of crises is a complex task that involves collecting accurate information and evaluating multiple data sources [Reinert et al., 2020]. Data visualization can support decision-makers and address problems related to the size and complexity of the data in such contexts [Dimara et al., 2021]. [Herschel and Clements, 2017] identify data storytelling as a topic in the field of data visualization that represents a structured approach to communicating relevant results from data analysis by combining data, visualization, and narratives. It can therefore be used to provide information on the data in context and to present the results of decisions (e.g., a lockdown, or reopening policies) to government actors.

Several studies describe best practices and quality criteria to create compelling visualizations [Nussbaumer Knaflic, 2015, Kosara and Mackinlay, 2013, Kosara, 2017, Segel and Heer, 2010, Tong et al., 2018b, Tong et al., 2018a, Boy et al., 2015, Nussbaumer Knaflic, 2012, Tufte, 1983]. However, research work on the benefits of implementing these practices in software development, or the drawbacks of not doing so, is still limited.

In this context, we present a case study based on the development of a quarantine information system for an Argentine region of about 1-million people. The system manages the monitoring of isolation and circulation permits during quarantine due to the COVID-19 pandemic. It had high volatility in its requirements and short development cycles, prioritizing the delivery of functionality and real-time visualizations for monitoring and decision making. This meant that visual and narrative aspects were set aside and that only some of the practices were implemented by the development team.

This work aims to determine the benefits of implementing data visualization best practices in the development of a software product, as well as the impact of not doing so. We argue that visualizations implementing the guidelines and best practices found in the literature would be more successful in conveying the information and thus be more effective in supporting the decision-making process.

We sought to achieve this goal by means of a case study. We analyzed the scientific literature and synthesized the data visualization best practices and quality criteria. We used quantitative and qualitative triangulation methods for the data collection and carried out questionnaires and experiments with system users whose role involved visualization-based decision making.

Our main contribution is a summary of data storytelling guidelines grouped into five categories, which can be used as a checklist and incorporated by teams in their development workflows when creating data visualizations. The selected guidelines are supported by initial empirical evidence, as provided by our case study.

The remainder of this paper is organized as follows. Section 2 presents the background and related works. Section 3 describes the research methodology, including the collection and analysis of the data. Section 4 presents the results, and Section 5 discusses the key findings. Finally, section 6 summarizes the main conclusions and addresses future work.

## 2 Background and Related Works

This work involves three main areas, namely: storytelling, data visualization, and visualization quality metrics, which are separately covered in the sub-sections below.

### 2.1 Data visualization in software applications

Data visualization has become essential for understanding large datasets and communicating findings. It provides a valuable instrument in cases where digital media enables assisted analysis [Kosara and Mackinlay, 2013]. Given the ubiquity of data, it is vital that visualizations can quickly and clearly expose intended patterns to audiences [Ajani et al., 2021].

As stated in [Gasser et al., 2020], data collection and usage are crucial in crisis response strategies, such as the COVID-19 pandemic. In this context, the rapid development of applications that improve data collection and further exploitation becomes necessary. For instance, these systems might need to be operational in a few months to support the decision-making of health or government actors. As a side-effect, these efforts might result in software systems with reduced quality or might affect the way of working of system developers [Ralph et al., 2020].

Additional examples of rapid application developments requiring data analysis for decision-making are [Chou et al., 2020] (less than a month) and [Da Silva, 2020] (two months). In the ELIS system, which was deployed in Bosnia and Herzegovina [Ponjavíc et al., 2020], one of its main objectives was simplicity and efficient communication for rapid decision-making. Along this line, [Perla et al., 2020] show the importance of visualizations and their evolution over time to meet the emerging needs of decision-making organizations. These studies agree that interactive maps and dashboards are useful tools for displaying key indicators for decision-making in crises. At last, [Dixit et al., 2020] present a tool developed in the context of the pandemic, in which the creation of visualizations was driven by a user-centric design process with prototypes and also by an iterative improvement of graphics. This process led to a set of steps to effectively create visualizations considering the needs of different participants (e.g., patients, health workers, or government authorities).

### 2.2 Storytelling and Data Visualization

A well-known definition of storytelling is presented by [Schreyögg, 2006] as: "the art of communicating ideas through stories." Data storytelling has gained increasing attention in the last few years, as it allows visualizations to effectively reveal information [Gershon and Page, 2001].

[Henry Riche et al., 2018] define "data-driven stories" as stories that are either based on or contain data, visualized to support one or more intended messages, usually including annotations (labels, pointers, text) or narration. [Kosara and Mackinlay,

2013] argue that data-driven storytelling is a natural next step for data analysis and visualization and a pivotal component for effective data exploration. Visualization dashboards are commonly used for decision-making but can be insufficient for communication purposes. In a recent study, [Sarikaya et al., 2019] found that people in Business Intelligence often put screenshots of dashboards into slide presentations, suggesting a need for more powerful storytelling features.

### 2.3 Data Storytelling Best Practices

Some studies provide examples of best practices for data visualization to improve the comprehension and clarity of data communications and achieve simple, logical stories. For instance, [Tufte, 1983] summarizes the basic principles of information visualization to convey complex ideas in a simple way. Similarly, in [Nussbaumer Knaflic, 2015] and [Nussbaumer Knaflic, 2012], the author describes the different types of visualizations, emphasizing the design strategies and narrative techniques that can be used to communicate clearly.

In [Segel and Heer, 2010], the authors discuss common design techniques in the media to create visual stories, as well as different genres and narrative structures. The article suggests that visualizations must strike a balance between a narrative flow intended by the author and a story discovery on the reader's part. In addition, [Kosara 2017] proposes a model to formalize the structure of data stories based on comparisons of patterns currently present in the media.

In [Tong et al., 2018b, Tong et al., 2018a], the authors present an overview of the most important elements in storytelling for data visualization and also describe the current challenges of the discipline. Furthermore, [Boy et al., 2015] investigates the benefits of interaction in information visualization to encourage user exploration and discusses how different interactive elements influence the level of user engagement. The best practices found in the literature are summarized in Table 1 and detailed in Section 3.4.

### 2.4 Quality Metrics in Visualizations

One of the main goals in data storytelling is to achieve effective, interesting visualizations by showing the most information in the simplest way possible [Behrisch et al., 2018], thus making the audience understand and remember the key points of the story [Henry Riche et al., 2018]. Different criteria can be used to measure the quality of visualizations, such as: engagement, comprehension, memorability, impact, dissemination, or credibility [Henry Riche et al., 2018]. For this study, we selected comprehension and memorability because they are essential in the decision-making process, as it is necessary to understand the data to extract useful information, find patterns and trends and develop business strategies. Both criteria are described below.

- **Comprehension:** It is defined as "reading and interpreting a graphic" [Friel, 2001]. Thus, this aspect is intrinsically related to practical knowledge or graphic literacy, that is, "the ability to read and interpret visually represented data and extract information from data visualizations" [Lee et al., 2017].
- **Memorability:** Brown et al. [Brown et al., 1977] define it as "the ability to maintain and retrieving information" that is related to the main goal of visualizations: to convey a message and to facilitate insight extraction [Henry Riche et al., 2018].

# 3 Research Methodology

This work pursues two goals: i) understand and analyze different data visualization strategies for software products, and ii) assess the impact of implementing those strategies (as best practices) on a decision-making process in environments with rapid development characteristics. In practical terms, we were interested in understanding how the usage of visualization best practices in a software product can influence the decision-making process while minimizing costs. To this end, we formulate the following research questions:

- **RQ1:** What storytelling best practices were met in the software product development?
- **RQ2:** What is the impact of those best practices that were not considered?
- **RQ3:** What are the benefits of implementing storytelling best practices?
- **RQ4:** Could storytelling best practices have been included during development without generating delays?

To answer the questions, we chose a case-study approach, as it allows researchers to understand the studied phenomenon and its context [Yin 2003]. We followed the steps proposed by [Yin 2003], namely: case-study design, data collection, evidence collection, analysis of collected data, and reports. Each of these steps is detailed in the following sub-sections.

## 3.1 System under Study

The case study involved an information system developed by a provincial government in Argentina during the COVID-19 pandemic. The system is intended to manage the monitoring of isolated persons and circulation permits in the context of the quarantine measures established in the country.

The system was built to provide historical and real-time data in the form of dashboards and interactive maps for decision-making. Because of the COVID-19 crisis, the system requirements undergo constant changes, and the system had to be developed in a short period of time, prioritizing the delivery of monitoring capabilities (e.g., charts). Due to these constraints, several visual and narrative aspects were disregarded in the first system version, and the development team addressed only a handful of best practices.

As part of the case study, we analyzed a total of 16 visualizations that monitor the progress of infections by reporting: active cases, total cumulative cases, people in isolation, as well as maps of focus of infection and population mobility rates. The remaining charts provided information about the circulation permits being processed through the system, such as: the number and type of permissions requested, their status, and locations with the most entries and departures. Figure 1 shows an instance of the visualizations used in the study.

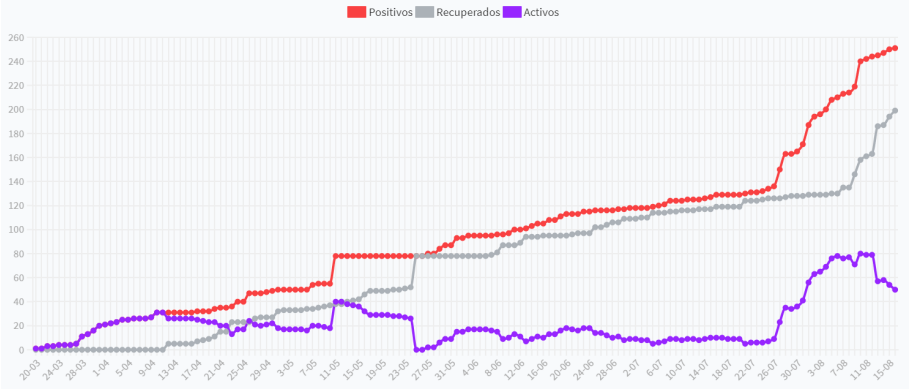


*Figure 1: Example of the visualizations analyzed in the study (before the improvements). The chart shows the new reported cases by day as well as the recovered and total cumulative cases.*

### 3.2 Study Design

The study was composed of four phases, in which the output of a given phase served as input for the next one, as shown in Figure **2**. The phases were the following:

- **Phase 1:** Analysis of the charts and improvement suggestions;
- **Phase 2:** Prioritization, estimation, planning, and implementation of the improvements;
- **Phase 3:** Evaluation of the quality of the charts after the improvements were made;
- **Phase 4:** Post-analysis of development times, when improvement tasks were included.

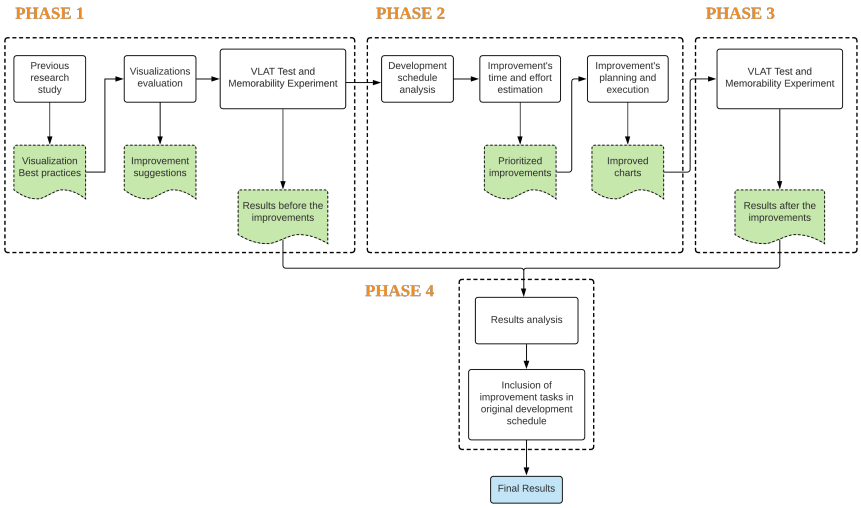


*Figure 2: Visual description of the research process*

### 3.3 Data Collection

We adopted a triangulation approach using quantitative and qualitative data for data collection since it allows us to combine different methodologies to achieve a complete and holistic view of the unit under study [Runeson and Höst, 2009].

**Phase 1:** Nine papers related to storytelling and best practices in data visualization were collected and studied [Nussbaumer Knaflic, 2015, Kosara and Mackinlay, 2013], [Kosara, 2017,Segel and Heer, 2010,Tong et al., 2018b,Tong et al., 2018a,Boy et al., 2015,Nussbaumer Knaflic, 2012,Tufte, 1983]. With the analysis results, a best practice guide was developed to allow researchers to evaluate visualizations and identify points of improvement.

To assess the comprehension of visualizations, we used the Visual Literacy Assessment Test (VLAT) [Lee et al., 2017]. In its original version, the test consists of 53 closed questions. For this study, we used only 19 items, taking into account the tasks associated with each type of chart: retrieve value; find extremum; determine range; find correlations/trends, and make comparisons [Lee et al., 2017].

Regarding memorability, we carried out an experiment like the one proposed by [Bateman et al., 2010] to learn about the impact of implementing visualization best practices. In particular, the test measures the degree of understanding along four axes:

- Subject: What is the chart about?
- Values: What are the displayed categories and values?
- Trend: Whether the chart shows any changes;
- Value Message: Whether the author is trying to communicate some message through the chart.

Based on the results of the evaluation of the charts, the VLAT responses, and the memorability experiment (ME), we obtained a number of possible improvements that were prioritized in Phase 2. Improvements included removing unnecessary elements like gridlines, data markers, and legends or changing the order in which the charts were presented. The complete list of improvements is presented in Section 3.4.

**Phase 2:** To estimate the priority and the mean execution time of the improvements, we applied the Wideband Delphi method, an estimation technique based on team consensus [Gandomani et al., 2014].

The prioritized improvements were presented to three software development teams for time estimation. In this case, the time range was unconstrained and expressed in minutes. Estimates were made on a five-point Likert scale ranging from 1 (not important) to 5 (very important).

**Phase 3:** Once the improvements planned in Phase 2 were implemented by the development team, we carried out the VLAT and the ME again to check for variations in information perception using best practices in data visualization. With both results, we applied comparative statistical analyses**.**

**Phase 4:** We audited the records and logs produced by the development team using a task tracking tool. Tasks devoted to product development and maintenance were analyzed to identify busy and spare time in the first 100 days of development. We also

inspected system-specific documentation created by the team to understand how the parts (of the product) affected by the improvements worked.

### 3.4 Data Analysis

The information collected in the four phases above was analyzed using different techniques.

Table 1 summarizes several best practices for storytelling and visualization, which were based on different articles from the literature. The first column of the table lists the primary sources of information. The guidelines were categorized into five groups, namely: Narrative, Design, Interaction, No Manipulation, and Appropriate Charts, which were subsequently divided into sub-categories. Narrative refers to how the story is told and includes: N1) the order of the event sequence, and N2) the incorporation of the basic elements of a story. Likewise, Design is made up of practices related to visualization itself, such as: D1) using consistent colors, D2) highlighting what is important, D3) eliminating clutter, and D4) using text, labels, and annotations to facilitate understanding. The Interaction category includes: I1) the incentive to explore and I2) the stimulation of user's curiosity. No Manipulation refers to ethics when creating visualizations. In this sense, it encompasses the practices: M1) not tampering with the data, M2) not citing out of context information, M3) changing the data, not the design, and M4) not distorting the charts. Finally, the Appropriate Graphics category aims at: A1) choosing a simple graphic, and A2) eliminating unnecessary complexity.

| | **Narrative** | | **Design** | | | | **Interaction** | | **No Manipulation** | | | | **Appropriate Chart** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Reference** | N1 | N2 | D1 | D2 | D3 | D4 | I1 | I2 | M1 | M2 | M3 | M4 | A1 | A2 |
| [Nussbaumer Knaflic, 2015] | | ✓ | ✓ | ✓ | ✓ | ✓ | | | ✓ | | ✓ | | ✓ | ✓ |
| [Kosara and Mackinlay, 2013] | | | | | | | ✓ | | | | | | | |
| [Kosara, 2017] | ✓ | | | | | | | | | | | | | |
| [Segel and Heer, 2010] | ✓ | | ✓ | | | ✓ | ✓ | | | | | | | |
| [Tong et al., 2018b] | ✓ | ✓ | | | | | | | | | | | | |
| [Tong et al., 2018a] | ✓ | ✓ | | | | | | | | | | | | |
| [Boy et al., 2015] | | | | | | | ✓ | | | | | | | |
| [Nussbaumer Knaflic, 2012] | | ✓ | ✓ | ✓ | ✓ | ✓ | | | ✓ | | ✓ | | ✓ | ✓ |
| [Tufte, 1983] | | | | | ✓ | ✓ | | ✓ | | ✓ | ✓ | ✓ | | ✓ |

*Table 1: Data storytelling best practices*

We assessed each of the charts present in the system to determine which practices were met. The results of this evaluation are summarized in Table 2.

Each chart was assigned an identification number, and a symbol was placed according to whether it complied (✓) with the practices of Table 1 (first column of Table 2). The row *%Total* indicates the total percentage of compliance for each practice in the analyzed charts. We observed that, while some elements of *Narrative* and *Design* could be improved, fundamental aspects such as the choice of a simple graph and the no manipulation of information are achieved in all graphics.

| | Narrative | | Design | | | | Interaction | | No Manipulation | | | | Appropriate Chart | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Chart Id.** | N1 | N2 | D1 | D2 | D3 | D4 | I1 | I2 | M1 | M2 | M3 | M4 | A1 | A2 |
| 1 | - | - | ✓ | - | - | - | - | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | - |
| 2 | - | - | ✓ | - | - | - | - | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | - |
| 3 | - | - | ✓ | - | - | - | - | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | - |
| 4 | - | - | ✓ | - | - | - | - | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | - |
| 5 | - | - | ✓ | - | - | - | - | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | - |
| 6 | - | - | ✓ | - | - | - | - | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | - |
| 7 | - | - | ✓ | - | - | - | - | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | - |
| 8 | - | - | ✓ | - | - | - | - | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | - |
| 9 | ✓ | - | ✓ | ✓ | - | - | - | - | ✓ | ✓ | ✓ | ✓ | ✓ | - |
| 10 | ✓ | ✓ | - | - | ✓ | ✓ | - | - | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 11 | ✓ | - | ✓ | ✓ | - | ✓ | - | - | ✓ | ✓ | ✓ | ✓ | ✓ | - |
| 12 | ✓ | ✓ | ✓ | - | - | ✓ | - | - | ✓ | ✓ | ✓ | ✓ | ✓ | - |
| 13 | ✓ | - | ✓ | - | - | ✓ | - | - | ✓ | ✓ | ✓ | ✓ | ✓ | - |
| 14 | - | - | - | ✓ | ✓ | ✓ | - | - | ✓ | ✓ | ✓ | ✓ | ✓ | - |
| 15 | ✓ | ✓ | - | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 16 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| % Total | 44 | 25 | 81 | 31 | 25 | 44 | 12 | 62 | 100 | 100 | 100 | 100 | 100 | 19 |

*Table 2: Results of the chart evaluation*

For both the VLAT and the ME, users with different roles and responsibilities were selected to achieve a global vision: health workers, administrative staff, IT staff (developers and technicians), and authorities.

**Visual Literacy Assessment Test**. The questionnaire was tested with 8 initial participants in order to correct or discard items if necessary. Participants were given a maximum time of 1 minute per question. The option to "skip" was included in each answer to avoid random responses. As a result, minor corrections were made to the questions to enhance the writing. The average execution time was 14 minutes.

For the execution of the questionnaire, 60 users were recruited, we ruled out 2 participants who were color-blind and other 6 participants whose answers were invalid, either because they responded randomly, taking a very short time (way below average) to complete the test, or did not complete it at all. A total of 52 participants remained. The response rate was 86%. From the 52 participants, 33 were women, and 19 were men, between 22 and 48 years old, with the mean age being 35 years.

As for the participants' roles, 14 were health workers, 17 were administrative staff, 12 were developers, and 9 were authorities. All participants are users of the system under study and use charts on a regular basis to make decisions.

The total possible score points on the test were 19 points. The scores for the participants ranged from 8 to 17 points ($M = 13{,}69$; $SD = 2{,}07$).

**Memorability Experiment**. The memorability experiment involved two phases. In the first phase (Description), the participants observed the charts for the time necessary to answer the four component questions. In the second phase (Recall and Recognition), the participants were asked to describe the aspects they remembered of the charts in as much detail as possible. The participants were assigned to one of two groups: immediate recall (15 minutes after the observation) and long-term recall (1 week after observation). The responses were recorded to facilitate further analysis and then scored according to the scale proposed in [Bateman et al., 2010]. The average time of each phase was 11 minutes, and the execution of the experiment took 22 minutes.

The experiment was conducted with 20 participants (13 women, 7 men) other than those who participated in the visual literacy test. From these 20 participants, 15% were experienced users using charts regularly, and the remaining 85% were health workers, administrative staff, and authorities who used charts only occasionally.

The responses were coded by a researcher and reviewed by a second researcher. In case of any difference, these were not greater than one point. In each phase, the total possible score points were 12 points. In the Description phase, the score ranged from 2 to 9 points ($M$ = 6,94; $SD$ = 1,77). For the Recall and Recognition phase, the scores ranged from 3 to 8 points for the immediate recall group ($M$ = 6,6; $SD$ = 1,65) and from 2 to 8 points for the long-term recall group ($M$ = 4.3; $SD$ = 2.26).

The results of the prioritization and estimation can be seen in Table 3. The *Source* column indicates the practice that originated the improvement suggestion during the evaluation and corresponds to those of Table 2 (N1, N2, D1 – D4, I1, A2). In ME1, the source was the memorability experiment, and the participants stated that it was not clear what information was being shown. Even though the VLAT was not a direct source, it helped us identify points of improvement. The *Chart Id* column identifies each of the visualizations present in the system. Column *P* indicates the priority (5 = high priority, 1 = low priority), and column *T* shows the estimated time for each improvement per chart expressed in minutes.

| **Source** | **Chart Id.** | **Improvement** | **P** | **T** |
|---|---|---|---|---|
| N1 | 1 - 8, 14 | Change the order of the charts. | 4,5 | 25 |
| N2 | 1 - 9, 11 | Add more descriptive titles. | 4,3 | 10 |
| N2 | 13 | Add bar chart. | 4,2 | 60 |
| N2 | 14 | Rearrange the elements. | 1,9 | 35 |
| D1 | 10, 14, 15 | Change colors. | 2,5 | 20 |
| D2 | 1 - 8 | Remove unnecessary data markers. | 3,2 | 25 |
| D2 | 10 | Color only what is most important. | 2,9 | 20 |
| D2 | 12, 13 | Remove pie charts. Replace them with bars. | 4,1 | 15 |
| D3 | 1 - 8 | Remove gridlines, avoid diagonal text, align elements to the left. | 2,3 | 10 |
| D3 | 9, 11 | Avoid mixing chart types; leave more white spaces. | 4,3 | 27 |
| D4 | 1 - 8 | Add information through pop-ups. | 3,8 | 25 |
| D4 | 9 | Add annotations. | 3,4 | 18 |
| I1 | 1 - 8 | Indicate the available interactions. E.g., "click to show or hide series." | 3,4 | 15 |
| A2 | 1 | Separate data into different charts. | 4,1 | 45 |
| A2 | 2 - 7 | Use the same chart type for all series. | 4,3 | 30 |
| ME1 | 8 | Explain what information is displayed on each axis. | 3,2 | 15 |

*Table 3: Estimation of time and priority per chart*

Once the improvements were completed and approved, we performed the visual literacy test and the memorability experiment again, according to the same steps described in Section 3.1. While the participants were the same in both phases, there was a 6-week time lapse between each iteration to avoid intentional learning. Following with the example presented in Figure 1, Figure 3 shows the improved visualization. The main changes involved: addition of explanatory titles and subtitles, removal of background and gridlines, and changes in the color of the data series. The data series were labeled directly instead of using legends, and the interactions (such as mouse hovering) were indicated explicitly. Information about the series was also added via pop-ups.

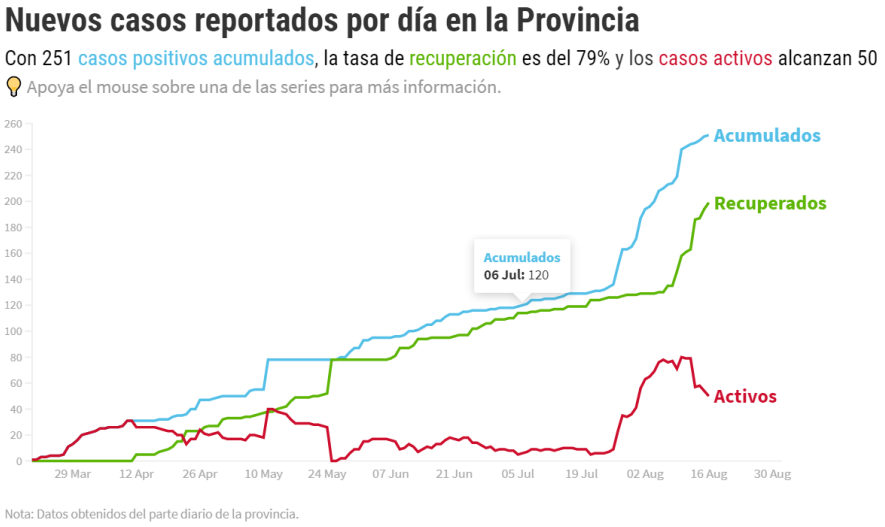


*Figure 3: Example of improved visualization, following the suggestions and best practices found in the literature*

For the visual literacy test, the questionnaire and the instructions given to the participants were the same as in Phase 1. In this case, the average execution time was 12.30 minutes, and the scores ranged from 10 to 19 points ($M$ = 15.33; $SD$ = 2.23). In some items, we observed a higher correct response rate (≥ 80%).

Regarding the memorability experiment, the steps described in the first iteration of the experiment were repeated. In this case, we observed that the participants required less time for the description and observation phase, which influenced the average execution time of the experiment, which was 18 minutes (4 minutes less than in Phase 1). The responses were again scored by one researcher and reviewed by another one. In the Description phase, the scores ranged from 4 to 10 points ($M$ =; $SD$ = 1,75). For the Recall and Recognition phase, the scores were 6 to 10 points ($M$ = 8.1; $SD$ = 1,29) for participants in the immediate recall group, and from 4 to 10 points ($M$ = 6.9; $SD$ = 2,6) for the long-term recall group.

Regarding the development schedule analysis, Figure 4 presents a summary of the distribution of work during the first 4 weeks of development, obtained from the records and logs kept by the team. The time used by the development team was divided into three possible states: resting times, working times, and timeouts or waiting times. The latter was the time spent between tasks waiting for approvals or new directives. The Y-

axis shows the hours of the day (0 - 23), and the colors green, red, and grey indicate resting times, working times, and timeouts, respectively.

The first few weeks (days 1 – 30), there were periods of intense work until the fundamental requirements were completed, followed by rather irregular breaks. The average daily working time was 10 hours, with 2-4 tasks per day. In cases where the time to implement new features was critical, the daily work exceeded the average, reaching up to 12 hours. Thus, in the following days, more resting times were observed. Waiting times or downtime in which other technical tasks were possible were employed to verify completed requirements, train health workers to use the system, or wait for new requirements or modifications to current functionalities. The average waiting time was 2.5 hours per day. Towards the third month (days 70 – 100), greater stability was achieved, with proportional working and resting times and lower waiting times.

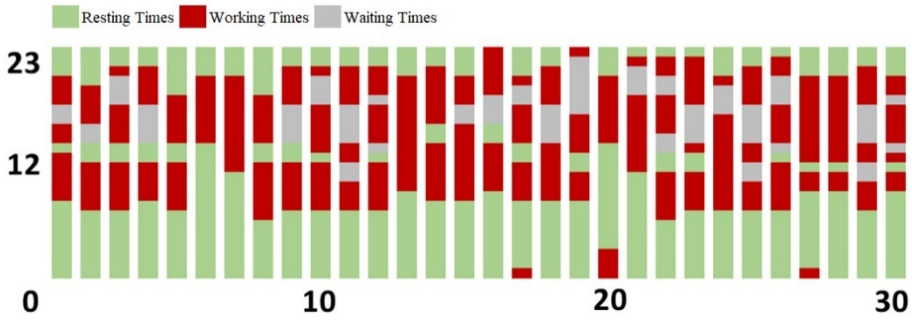


*Figure 4: Times used for development tasks during the first month*

Table 4 summarizes significant waiting times per week. For this study, we considered a significant time of 60 minutes, given that a developer needs a time of understanding, preparation, construction, and testing of a requirement [Jørgensen, 2004].

| Week | Waiting Times | Week | Waiting Times |
|---|---|---|---|
| 1 | 11 h. | 9 | 4 h. |
| 2 | 15 h. | 10 | 2 h. |
| 3 | 21 h. | 11 | - |
| 4 | 21 h. | 12 | 15 h. |
| 5 | 8 h. | 13 | 10 h. |
| 6 | 4 h | 14 | - |
| 7 | 6 h | 15 | 5 h. |
| 8 | 8 h. | | |

*Table 4: Significant waiting times during the software development*

## 4 Results

This section presents the findings of the study. We discuss the results for the VLAT and the ME before and after improvements were made to visualizations. We also comment on the inclusion of the improvement tasks in the original development schedule.

**Visual Literacy Assessment Test.** In Phase 1, the participants scored an average of 13.69 points (SD = 2.07), with scores ranging from 8 to 17 points. In Phase 3, the average score was 15,33 (SD = 2.23), with a minimum of 10 points and a maximum of 19. In addition, we performed a Shapiro-Wilk test [Royston, 1982] to check the normality of the scores and facilitate observations. The results showed that scores in both phases follow a normal distribution – ($w = 0.956$, $p = 0.053$) and ($w = 0.956$, $p = 0.057$) for the first and third phases, respectively.

At the item level, we conducted a McNemar test [Eliasziw and Donner, 1991] for paired proportions with a significance level $\alpha = 0.001$, under the hypothesis that best practices and improvements would increase comprehension. Figure 5 shows the results. The asterisks (*) indicate the items in which a significant difference in comprehension was observed. According to these results, significant differences were found in 14 of the 19 items. Among them, 3 items (items 10, 15, 19) had high scores on both occasions (the correct answer rate was equal to or greater than 80%). For the remaining items that had a significant difference, 4 of them were associated with the task of Identify Values (items 1, 16, 13, 17), 2 with Find Extremums (items 11, 14), 2 with Determine Ranges (items 3, 8), 2 with Make Comparisons (items 5, 12), and 1 with Identify Trends (item 16).

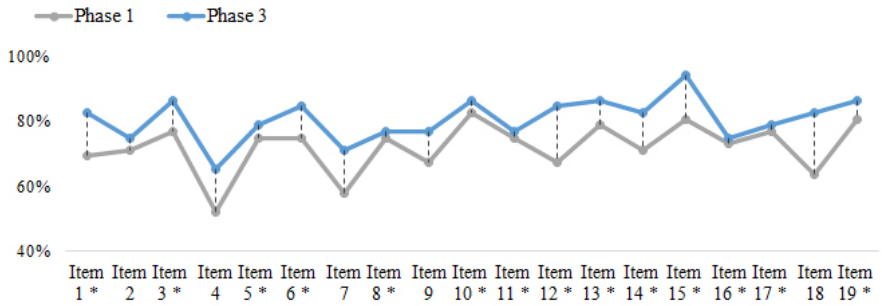


*Figure 5: Percentage of correct answers per item in each iteration of the Visual Literacy Assessment Test*

The overall performance of the participants was also analyzed. They achieved relatively high scores in both cases: the correct answer rate was more than 40%. Some participants maintained the same scores in phases 1 and 3. However, there were cases (participants 15, 27, 45) in which we observed a considerable increase in the scores (see Figure 6). In addition, we performed a pairwise t-test to determine whether there was a significant difference between the score means in both iterations of the test, with $\alpha = 0.05$. The

result was $p = 6{,}39513E-12$, which was sufficient evidence to reject the null hypothesis and infer that the improvements increased graph comprehension.

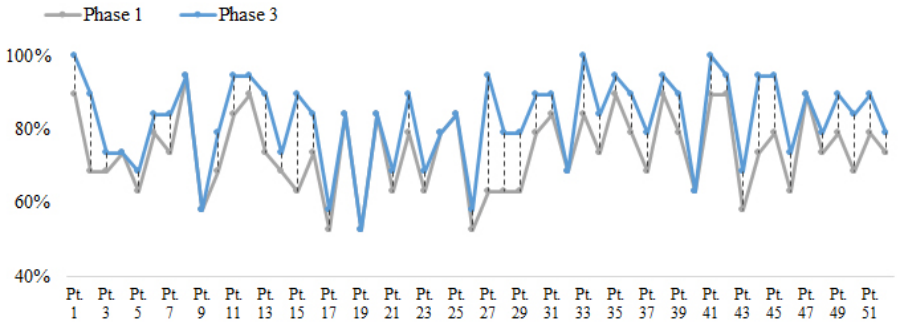


*Figure 6: Participants' performance in each iteration of the Visual Literacy Assessment Test*

**Memorability experiment.** The scores in both the Description and Recall and Recognition phases were computed based on the coded responses. We applied a one-tailed pairwise t-test with a significance level $\alpha = 0.05$ to determine whether the presence of visualization best practices altered the quality of participant's chart descriptions. The null hypothesis was that best practices would aid the interpretation of charts. These results are shown graphically in Figure 7. Asterisks (*) indicate significant differences. There were no significant differences in the Description phase for the chart subject ($t = -0.44$, $p = 0.33$), or the trend ($t = 0$, $p = 0.5$). However, we observed differences for the values ($t = -1.8$, $p = 0.04$) and the message ($t = -1.75$, $p = 0.04$), which were easier to identify with the improved charts, as seen in Figure 7.

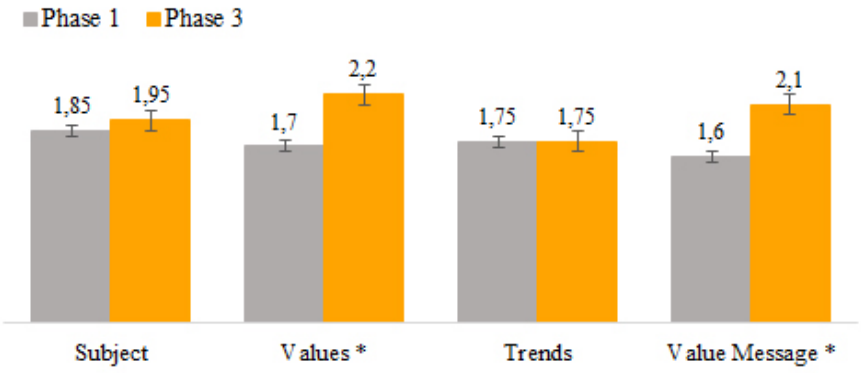


*Figure 7: Mean ± Standard Deviation of scores in the Description phase of the memorability experiment*

As for the Recall and Recognition phase, participants in the short-term group showed no significant differences for the chart subject ($t = -0.69$, $p = 0.25$), the trend ($t = -0.43$, $p = 0.34$), or the message ($t = -0.64$, $p = 0.27$). However, there was a difference in the chart values ($t = -1.75$, $p = 0.03$). For participants in the long-term recall group, there were significant differences in the subject ($t = -3.77$, $p = 0.002$) and the values ($t = -1.77$, $p = 0.05$) of the chart; we observed no differences in the trend ($t = -0.43$, $p = 0.34$) and the message ($t = -0.58$, $p = 0.29$), as shown in Figure 8.

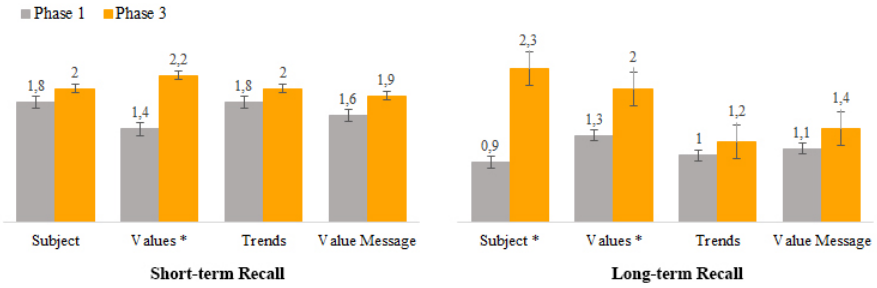


*Figure 8: Mean ± Standard Deviation of scores in the Recall and Recognition phase of the memorability experiment*

We also observed participants' performance across the two iterations of the experiment (see Figure 9). In the first one, the score for the Description phase ranged from 2 to 9 points ($M = 6.9$; $SD = -1.77$). For the Recall and Recognition phase, participants in the short-term group scored between 3 and 8 points ($M = 6.6$; $SD = 1.65$), while those in the long-term group scored between 2 and 8 points ($M = 4.3$; $SD = 2.26$). In the second iteration, the score ranged from 4 to 10 points for the Description phase ($M = 8$; $SD = 1.75$). For the Recognition and Recall phase, participants in the short-term group scored between 6 and 10 points ($M = 8.1$; $SD = 1.29$) while the long-term group scored between 4 and 10 points ($M = 6.9$; $SD = 2.6$).

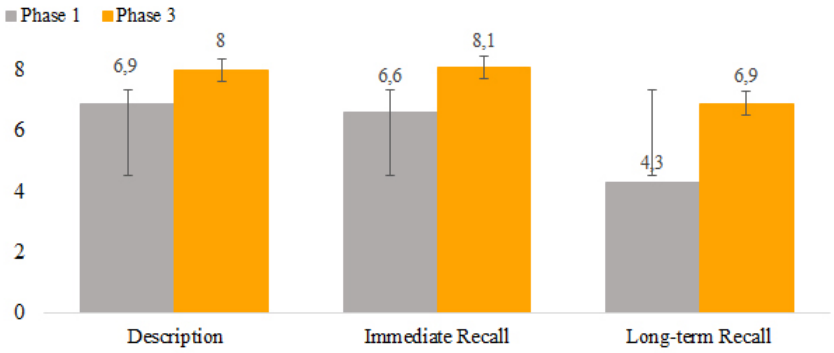


*Figure 9: Mean ± Standard Deviation of scores in Description and Recall and Recognition phases in each iteration of the memorability experiment*

Table 5 presents the actual development times recorded in the improvement tasks. In particular, the *TDT* column indicates the total development time of each improvement, taking into account all charts expressed in minutes.

| Source | Chart Id. | Improvement | P | TDT |
|---|---|---|---|---|
| N1 | 1 - 8, 14 | Change the order of the charts. | 1,9 | 40 |
| N2 | 1 - 9, 11 | Add more descriptive titles. | 2,5 | 45 |
| N2 | 13 | Add bar chart. | 3,2 | 105 |
| N2 | 14 | Re-arrange the elements. | 2,9 | 30 |
| D1 | 10, 14, 15 | Change colors. | 4,1 | 20 |
| D2 | 1 - 8 | Remove unnecessary data markers. | 2,3 | 75 |
| D2 | 10 | Color only what is most important. | 4,3 | 45 |
| D2 | 12, 13 | Remove pie charts. Replace them with bars. | 3,8 | 150 |
| D3 | 1 - 8 | Remove gridlines, avoid diagonal text, align elements to the left. | 3,4 | 30 |
| D3 | 9, 11 | Avoid mixing chart types; leave more white spaces. | 3,4 | 120 |
| D4 | 1 - 8 | Add information through pop-ups. | 4,1 | 90 |
| D4 | 9 | Add annotations. | 4,3 | 80 |
| I1 | 1 - 8 | Indicate the available interactions. E.g., "click to show or hide series." | 3,2 | 45 |
| A2 | 1 | Separate data into different charts. | 4,1 | 45 |
| A2 | 2 - 7 | Use the same chart type for all series. | 4,3 | 30 |
| E1 | 8 | Explain what information is displayed on each axis. | 3,2 | 15 |

*Table 5: Summary of total development times per improvement*

Figure 10 shows the development times having incorporated the improvement tasks (shown in yellow) into the timeouts recorded in Phase 4 of section 3.1. The analysis took into account waiting times and the date on which the development team originally carried out the construction of the chart. We observed that the improvements would not have taken extra time if they were included from the beginning of the development process.

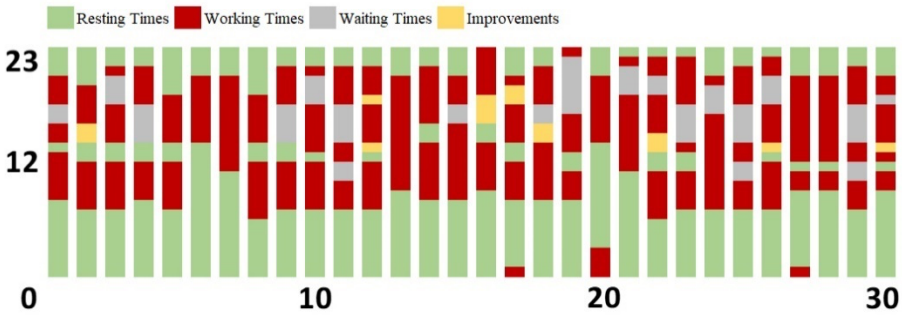


*Figure 10: Some improvements could have been implemented during waiting times generated by other technical tasks*

## 5 Discussion

This section discusses the results obtained and the answers to our research questions.

**RQ1: What storytelling best practices were met in the software product development?** As Table 2 shows, the practices related to *Narrative*, *Design,* and *Interaction* were the least fulfilled in the creation of the charts. On the other hand, those practices related to *No Manipulation of Information* and *Appropriate Charts* were present in most cases.

We present below the details of each practice and its percentage of compliance.

- *Narrative.* This feature was affected mainly by the absence of explanatory titles, which does not provide a clear idea of what information is presented on each chart, coupled with the inconvenient order in which the charts were arranged.

- *Design.* Some factors contributing negatively to the design were: lack of white space, diagonally oriented data labels, and an abundance of data markers. Also, some milestones in the data could have been highlighted by using text and annotations.

- *Interaction.* Except for the charts monitoring infections (1 – 8), most charts did not offer any kind of interaction for the user. In cases in which there were interactive elements, they were not explicitly indicated, and users had to find them on their own, losing the opportunity for the "dialogue" between the user and the data, as [Dimara and Perin, 2020] point out.

- *No Manipulation.* Aspects of fidelity and no manipulation of information, which are key factors when communicating, were fulfilled in all the charts evaluated. The information was accurate and not biased.

- *Appropriate Charts*. In most cases, simple charts were chosen to represent the information. However, there were cases in which the reading and interpretation of the data could have been further facilitated. For instance, pie charts could have been replaced by horizontal bar charts.

**RQ2: What is the impact of those best practices that were not considered?** The results of the VLAT (in Phase 1) indicate that users had greater difficulty answering questions when the chart elements were not clear enough. This situation is evident in items 1 and 12, which had the lowest rate of correct answers (less than 70%). These items were associated with the Identify Value and Make Comparisons tasks, respectively.

As for the memorability experiment, some users required more time to answer questions in the Observation and Description phase. For the Recall and Recognition phase, participants located in the short-term group had no problems describing the details of the charts. Those participants with more experience in data visualizations suggested possible improvements to aid interpretation. Participants in the long-term recall group, on the other hand, required more suggestions to remember the fine details of the charts.

**RQ3: What are the benefits of implementing storytelling best practices?** The results of both the VLAT and the ME (in Phase 3) show a positive relationship between the implementation of best practices in data visualization and the ability to understand and remember graphs by users. Users scored significantly higher than in the previous iteration (carried out in Phase 1).

In the VLAT, users performed better with the tasks of Identifying Values, Finding Extremes, Determining Ranges, Making Comparisons, and Identifying Trends.

In the memorability experiment, there were significant differences in the identification of the subject and values of the chart, but not for the cases of Trends and Message. This fact indicates that visualizations incorporating narrative strategies seem to generate user interaction and eliminate visual clutter, conveying information more clearly to users. This is consistent with [Tufte, 1983], which proposes the data-to-ink ratio, where the author argues that any ink that is not used to present data must be removed. Similarly, [Nussbaumer Knaflic, 2015] offers some recommendations to reduce the effort required to interpret the information behind a chart.

Overall, these findings confirmed the importance of data storytelling as means to facilitate the understanding and memorization of the information [Shi et al., 2020].

**RQ4: Could storytelling best practices have been included during development without generating delays?** The developers often neglected the design of the graphics in favor of other aspects and functionalities of the system due to the lack of knowledge about good practices in data visualization. After observing the system development logs, we confirmed that poor visualizations were not due to the lack of time but rather to the lack of knowledge and guidance when creating visualizations. In particular, narrative-related tasks (e.g., changing the order of graphics or adding descriptive titles) and design (removing unnecessary data markers, removing grid lines, or adding annotations) could have been performed during timeouts, as they did not represent

major challenges for the team. This observation highlights the need to institutionalize the guidelines so they can be addressed from the beginning of the development process.

### 5.1 Threats to Validity

Validity refers to the reliability of the results, i.e., the extent to which the results are valid and not influenced by the perspective of the researchers. We considered the four aspects of validity as proposed in [Runeson and Höst, 2009].

**Construct validity.** It reflects the extent to which the research methodology represents what the researchers have in mind and what is investigated in relation to the research questions [Runeson and Höst, 2009]. In our study, the quality criteria considered are subjective, and as such, their definition depends on the context, and they are not directly measurable [Henry Riche et al., 2018]. To evaluate the criteria, we used previously validated instruments [Lee et al., 2017, Bateman et al., 2010] and had other researchers verify them.

**Internal validity.** It analyzes risks when studying causal relationships. When the researcher studies whether a factor affects an investigated factor, there is a risk that the latter will also be affected by a third factor [Runeson and Höst, 2009]. If the researcher is unaware of the third factor or its degree of interference, there is a threat to internal validity. The VLAT and the ME were conducted with the same participants, which can be a risk of intentional learning that could skew the results. This threat was mitigated with a time span of six weeks between each iteration of the experiments.

**External validity.** It refers to what extent it is possible to generalize findings beyond the case under study and to what extent they are of interest to the public outside the organization where the research is carried out. While this research was conducted on a single information system, the purpose of these types of studies is to enable generalization when the results extend to cases that have common characteristics, and therefore for which the findings are relevant [Runeson and Höst, 2009]. To mitigate this threat, users with different roles, responsibilities, and experience were selected for both the VLAT and the ME.

**Reliability.** It shows the extent to which the research data and findings are independent of researchers. That is, if another author conducted the same study, the results should be the same or similar [Runeson and Höst, 2009]. This threat was mitigated by having three researchers conducting the study. In addition, the steps of this research, together with the instruments used, the questionnaire, and the experiment, were all reviewed by a third author.

## 6 Conclusions

This article presents the results of a case study that investigates the role of data visualization best practices in the development of a software product, as well as the benefits of implementing them and the impact of those practices that were not included. The data were collected through the study of 9 articles from the literature, the execution of 52 questionnaires, 20 experiments, and the analysis of the system development tasks log.

We found that the usage of storytelling techniques aids the decision-making process by implementing narrative and design strategies that increase understanding

and memorability of visualizations, thus facilitating the process of interpreting data, finding patterns, and extracting information for the system stakeholders.

Among the reasons for not applying visualization practices by the team, we noticed that they seemed to be the lack of knowledge and poor preparation in terms of information visualization and storytelling techniques, rather than the lack of time, considering that the practices often represent small tasks and can be performed during waiting times. In the absence of these skills, developers rely on the default settings of tools, making stories lose their potential or become difficult to understand. Therefore, to identify improvement points, we argue that teams need to know the context and needs for each case, properly analyze the graphs, and learn about visualization best practices.

This work introduces a series of guidelines grouped into five categories that can be used as a checklist and incorporated by teams in their development workflows when creating data-driven visualizations. The guidelines also provide insights into how the implementation of these practices influence comprehension and memorability of charts, which fosters better decision-making. We suggest integrating some kind of monitoring to minimize deviations when new charts are added, or the current ones need to be modified.

Finally, the results of this study provide a strong basis for future research. Future work includes the execution of a secondary mapping study to review the guidelines and quality criteria associated with data-driven storytelling, as well as the development of a model to evaluate the quality of visualizations.

**Acknowledgments**

This work was supported by the National University of the NorthEast (SCyT - UNNE) under grants 17F017 and 17F018. The authors wish to thank Gladys Dapozo for the constructive feedback.

## References

[Ajani et al., 2021] K. Ajani, E. Lee, C. Xiong, C. Nussbaumer Knaflic, W. Kemper and S. Franconeri, "Declutter and Focus: Empirically Evaluating Design Guidelines for Effective Data Communication," in IEEE Transactions on Visualization and Computer Graphics, doi: 10.1109/TVCG.2021.3068337.

[Baker, 2017] Baker, R. (2017). Agile UX Storytelling. Apress, Berkeley, CA.

[Bateman et al., 2010] Bateman, S., Mandryk, R. L., Gutwin, C., Genest, A., McDine, D., and Brooks, C. (2010). Useful junk? The effects of visual embellishment on comprehension and memorability of charts. Conference on Human Factors in Computing Systems - Proceedings, 4:2573–2582.

[Behrisch et al., 2018] Behrisch, M., Blumenschein, M., Kim, N. W., Shao, L., ElAssady, M., Fuchs, J., Seebacher, D., Diehl, A., Brandes, U., Pfister, H., Schreck, T., Weiskopf, D., and Keim, D. A. (2018). Quality Metrics for Information Visualization. Computer Graphics Forum, 37(3):625–662.

[Boy et al., 2015] Boy, J., Detienne, F., and Fekete, J. D. (2015). Storytelling in information visualizations: Does it engage users to explore data? Conference on Human Factors in Computing Systems - Proceedings, 2015-April:1449–1458.

[Brown et al., 1977] Brown, J., Lewis, V. J., and Monk, A. F. (1977). Memorability, Word Frequency, and Negative Recognition. Quarterly Journal of Experimental Psychology, 29(3):461–473.

[Chou et al., 2020] Chou, S.-H., Kearns, J., Turk, P., Kowalkowski, M., Roberge, J., Priem, J., Taylor, Y., Burns, R., Palmer, P., and McWilliams, A. (2020). COVID19 Utilization and Resource Visualization Engine (CURVE) to Forecast In-Hospital Resources. medRxiv, page 2020.05.01.20087973.

[Cohen et al., 1996] Cohen, M. A., Eliashberg, J., and Ho, T. H. (1996). New product development: The performance and time-to-market tradeoff. Management Science, 42(2):173–186.

[Da Silva, 2020] Da Silva, C. M. N. G. (2020). Comparative Analysis of Business Intelligence Resulting from the Brazilian Ministry of Health Database of COVID-19 Contamination.

[Dimara and Perin, 2020] Dimara, Evanthia & Perin, Charles. (2020). What is Interaction for Data Visualization?. IEEE Transactions on Visualization and Computer Graphics.

[Dimara et al., 2021] E. Dimara, H. Zhang, M. Tory and S. Franconeri, "The Unmet Data Visualization Needs of Decision Makers within Organizations," in IEEE Transactions on Visualization and Computer Graphics, doi: 10.1109/TVCG.2021.3074023.

[Dixit et al., 2020] Dixit, R. A., Hurst, S., Adams, K. T., Boxley, C., LysenHendershot, K., Bennett, S. S., Booker, E., and Ratwani, R. M. (2020). Rapid development of visualization dashboards to enhance situation awareness of COVID-19 telehealth initiatives at a multihospital healthcare system. Journal of the American Medical Informatics Association: JAMIA, 27(9):1456–1461.

[Eliasziw and Donner, 1991] Eliasziw, M. and Donner, A. (1991). Application of the McNemar test to non-independent matched pair data. Statistics in Medicine, 10(12):1981–1991.

[Friel, 2001] Friel, S. N. (2001). Making Sense of Graphs: Critical Factors Influencing Comprehension and Instructional Implications. Technical Report 2.

[Gandomani et al., 2014] Gandomani, T. J., Wei, K. T., and Binhamid, A. K. (2014). A case study research on software cost estimation using experts' estimates, Wideband Delphi, and Planning Poker technique. International Journal of Software Engineering and its Applications, 8(11):173–182.

[Gasser et al., 2020] Gasser, U., Ienca, M., Scheibner, J., Sleigh, J., and Vayena, E. (2020). Digital tools against COVID-19: taxonomy, ethical challenges, and navigation aid.

[Henry Riche et al., 2018] Henry Riche, N., Hurter, C., Diakopoulos, N., and Carpendale, S. (2018). Data-Driven Storytelling.

[Herschel, R. and Clements, N. 2017] Herschel, R., & Clements, N. (2017). The Importance of Storytelling in Business Intelligence. Int. J. Bus. Intell. Res., 8, 26-39.

[Janssen and van der Voort, 2020] Janssen, M. and van der Voort, H. (2020). Agile and adaptive governance in crisis response: Lessons from the COVID-19 pandemic. International Journal of Information Management, 55:102180.

[Jørgensen, 2004] Jørgensen, M. (2004). A review of studies on expert estimation of software development effort. J. Syst. Softw., 70(1-2):37–60.

[Kosara and Mackinlay, 2013] Kosara, R. and Mackinlay, J. (2013). Storytelling: The Next Step for Visualization.

[Kosara, 2017] Kosara, R. (2017). An Argument Structure for Data Stories. Technical report.

[Krausz et al., 2020] Krausz, M., Westenberg, J. N., Vigo, D., Spence, R. T., and Ramsey, D. (2020). Emergency Response to COVID-19 in Canada: Platform Development and Implementation for eHealth in Crisis Management. *JMIR Public Health and*

[Lee et al., 2017] Lee, S., Kim, S. H., and Kwon, B. C. (2017). VLAT: Development of a Visualization Literacy Assessment Test. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):551–560.

[Nussbaumer Knaflic, 2012] Nussbaumer Knaflic, C. (2012). Data Stories. Technical report.

[Nussbaumer Knaflic, 2015] Nussbaumer Knaflic, C. (2015). *Storytelling With Data: A data visualization guide for bussiness professionals*. John Wiley & Sons, Ltd., Hoboken, New Jersey.

[Perla et al., 2020] Perla, R. J., Provost, S. M., Parry, G. J., Little, K., and Provost, L. P. (2020). Understanding variation in reported covid-19 deaths with a novel Shewhart chart application. *International Journal for Quality in Health Care*, 2020:1–8.

[Ponjavíc et al., 2020] Ponjavíc, M., Karabegovíc, A., Ferhatbegovíc, E., Tahirovíc, E., Uzunovíc, S., Travar, M., Pilav, A., Mulíc, M., Karakaˇs, S., Avdíc, N., Mulabdíc, Z., Pavíc, G., Biˇco, M., Vasilj, I., Mami´c, D., and Huki´c, M. (2020). Spatio-temporal data visualization for monitoring of control measures in the prevention of the spread of COVID-19 in Bosnia and Herzegovina. *Medicinski Glasnik*, 17(2):1–10.

[Rahman and Farhana, 2020] Rahman, A. and Farhana, E. (2020). An Exploratory Characterization of Bugs in COVID-19 Software Projects.

[Ralph et al., 2020] Ralph, P., Baltes, S., Adisaputri, G., Torkar, R., Kovalenko, V., Kalinowski, M., Novielli, N., Yoo, S., Devroey, X., Tan, X., Zhou, M., Turhan, B., Hoda, R., Hata, H., Robles, G., Fard, A. M., and Alkadhi, R. (2020). Pandemic Programming: How COVID-19 affects software developers and how their organizations can help. *Empirical Software Engineering*, 25(6):4927–4961.

[Reinert et al., 2020] A. Reinert et al., "Visual Analytics for Decision-Making During Pandemics," in Computing in Science & Engineering, vol. 22, no. 6, pp. 48-59, 1 Nov.-Dec. 2020, doi: 10.1109/MCSE.2020.3023288.

[Royston, 1982] Royston, J. P. (1982). An Extension of Shapiro and Wilk's W Test for Normality to Large Samples. *Applied Statistics*, 31(2):115.

[Runeson and Höst, 2009] Runeson, P. and Höst, M. (2009). Guidelines for conducting and reporting case study research in software engineering. *Empirical Software Engineering*, 14(2):131–164.

[Sarikaya et al., 2019] A. Sarikaya, M. Correll, L. Bartram, M. Tory, and D. Fisher, "What Do We Talk About When We Talk About Dashboards?," in IEEE Transactions on Visualization and Computer Graphics, vol. 25, no. 1, pp. 682-692, Jan. 2019, doi: 10.1109/TVCG.2018.2864903.

[Schreyögg, 2006] Schreyögg, G. (2006). *Knowledge Management and Narratives: Organizational Effectiveness Through Storytelling*.

[Segel and Heer, 2010] Segel, E. and Heer, J. (2010). Narrative Visualization: Telling Stories with Data.

[Shi et al., 2020] Shi, Danqing & Xu, Xinyue & Sun, Fuling & Shi, Yang & Cao, Nan. (2020). Calliope: Automatic Visual Data Story Generation from a Spreadsheet.

*Surveillance*, 6(2):e18995.

[Tong et al., 2018a] Tong, C., Roberts, R., Borgo, R., Walton, S., Laramee, R. S., Wegba, K., Lu, A., Wang, Y., Qu, H., Luo, Q., and Ma, X. (2018a). Storytelling and visualization: An extended survey. *Information (Switzerland)*, 9(3). [Tong et al., 2018b] Tong, C., Roberts, R., Laramee, R. S., Wegba, K., Lu, A., Wang, Y., Qu, H., Luo, Q., and Ma, X. (2018b). Storytelling and visualization: A survey.

[Tufte, 1983] Tufte, E. R. (1983). *The Visual Display of Quantitative Information*.

*VISIGRAPP 2018 - Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 3:212–224.

[Yin, 2003] Yin, R. K. (2003). *Case Study Research Design and Methods*.

[Zimmerman, 1997] Zimmerman, B. B. (1997). Applying Tufte's principles of information design to creating effective Web sites. *ACM SIGDOC Annual International Conference on Computer Documentation, Proceedings*, pages 309–317.