Developing a Visualization Application

A possible definition of Visualization is “the representation and presentation of data exploiting our visual perception capabilities to amplify cognition” (Kirk, 2012). This definition encompasses two fundamental phases in the process of designing a visualization: the decision on how the data will be visually encoded (represented) through visual structures organized in a certain way (visualization techniques, which can be, for example, a bar graph or a scatter plot); and the decision on how, where and when these representations will be presented to the user, to produce the images to be displayed corresponding to various views, always in order to provide a greater understanding of the data and, ultimately, of the phenomenon that gave rise to them. Although these two phases are very important, before anything else, it is essential to understand the problem domain, what should be shown, to whom and why.

Although initially the visualizations were static, with the representation and presentation being fixed, nowadays situations in which the aim is to allow visual exploration of the data are very common, so it will be necessary to provide for the possibility of interaction on the part of the user (for example, allowing filtering the data they want to visualize, choose the visual representations to use or change the views produced).

Effectively, any visualization is intended to be seen by humans. It does not have to generate the insights, calculate the values, or present the conclusions. Visualization makes information more evident and understandable, but it is the human with their knowledge and their questions, in their context, who reaches the conclusions. It is for this reason that we have to design visualization taking all of this into account, to properly support the user, and it is essential to use a Human-Centered Design[1] methodology to develop visualization solutions that can help achieve the user´s goals.

This fundamental human involvement makes visualization a very different approach to machine learning, although both can be used to solve the same problems (possibly at different moments in a data analysis process).

The process of developing a visualization application includes numerous choices, namely the choice of techniques to be used in each phase, most appropriate to the specific problem in question. Like any design process, it is arduous, very rarely simple and linear, and has to meet a series of constraints specific to the problem at hand. Thinking that there is a simple set of rules that guarantee an effective solution is idealistic. However, there is already knowledge, whether theoretical or based on practice, that allows us to understand which techniques work best for each situation. There are also development approaches that increase the likelihood of achieving an effective solution. Methodologies of the type generally referred to as Human-Centered Design have been used successfully to develop visualization solutions. They are iterative and begin by considering the needs and characteristics of the users for whom the visualization is being developed. They should include iterations with implementation of prototypes of increasing fidelity, evaluation and redesign (as illustrated in Figure 1), until a solution is reached that can meet the objectives.

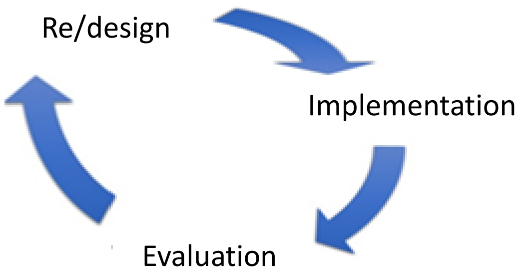

Figure 1. Iterative approach to develop a Visualization application.

This design process is part of a larger visualization production process, which must be carried out, as previously mentioned, using a methodology that takes into consideration the problem domain, the characteristics of the users for whom the visualizations are intended, the tasks

[1] https://www.interaction-design.org/literature/topics/human-centered-design?

that these will have to perform with the help of visualizations and the context in which they will do so, but also the access and cleaning of the data to be represented, as well as technical aspects of implementation.

However, if these are the operations necessary to produce a visualization (treating the data, visually coding it, producing the images and allowing the user to intervene throughout the process), the different and fundamental question arises: "how to inform all the decisions that must be made throughout all these phases to be able to produce a solution that supports users in the best way possible?" This is achieved through an appropriate development methodology. Creating a visualization is not an exact science. There will rarely, if ever, be a single right answer or optimal solution; it is much more about a design process to arrive at solutions that are satisfactory.

It should be noted that there are two possible "angles of attack" for visualization design: **top down** or **bottom up**[2]. Problem-driven work, the most common approach, starts at the top domain situation level and work down through abstraction, visualization technique, and algorithms. In technique-driven work, the work starts at one of the bottom two levels, visualization technique or algorithm design, inventing new techniques, or new algorithms to implement existing visualization techniques.

Problem-driven visualization, starts by trying to understand the problems of some real-world users and designing a solution that helps them work more effectively. In visualization literature, this is generally called a design study. Most often the problem can be solved using existing visualization techniques and the challenge lies at the abstraction level. While sometimes the problem motivates the creation of new techniques, if no existing ones seem to adequately solve the problem, this is an endeavor more often conducted in the scope of research and not development. In what follows several important aspects of a methodology for developing visualization solutions using the top-down approach are addressed: the design process stages, a taxonomy of abstract tasks, as well as ways to evaluate and validate the choice of techniques to use. All this is vital to ensure, as much as possible, that a useful solution for the intended users is obtained.

[2] https://learning.oreilly.com/library/view/visualization-analysis-and/9781466508910/K14708_C004.xhtml#c04_sec4-4

1. Design Process Stages

The process of designing a visualization application (or just a visualization) is an iterative process (the various phases are repeated and refined until a good solution is reached), centered on the user (the user has a central and participatory aspect in the design process). The four main stages of this methodology allow the designer to understand the relevant aspects of the process, may be summarized as four key questions, extending the methodology proposed by (Munzner, 2014):

1. Who are the users and in what context do they operate (**who?**).

2. Their tasks and questions they want to answer, i.e. why the visualization is needed. (**why?)**

3. Obtain/process/understand the data available for this purpose. (**what?)**

4. The best way to visually represent this data, in order to allow users to carry out their tasks and answer their questions. (**how?**)

This visualization design process also has points of contact with the so-called **Information Visualization reference model**, presented in (Card et al., 1999). This, simpler one, focuses in particular on the last two phases, but already highlights the notion that the user is part of the process, which gives it the **human-in-the-loop** nature of the process (Figure 2). Without the human this process makes no sense, and it is desirable to provide the user with the possibility of intervening in all phases through interaction mechanisms as indicated by the arrows that connect the user to each of the phases. It is also desirable that they can obtain other data, if the already available data are not sufficient nor adequate to carry out the intended tasks.

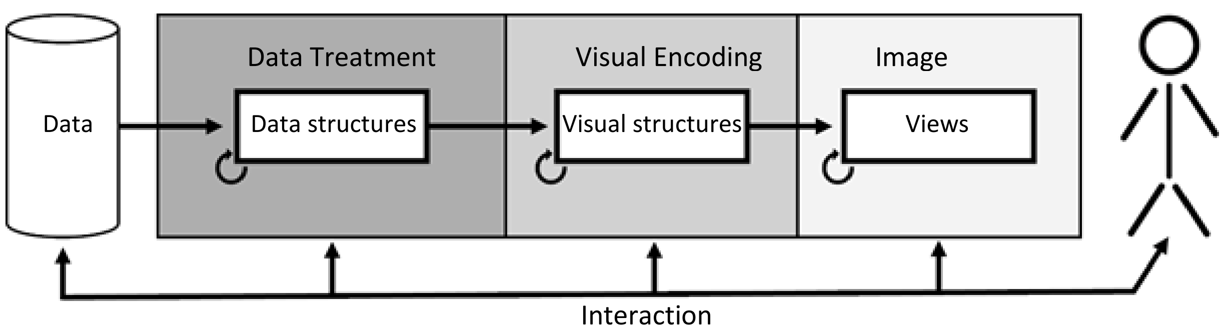

Figure 2. The InfoVis reference model (adapted from Card et al., 1999), plus the possibility of getting more data.

1.1 Who: Users and Context

The first aspect to understand is **who** the visualization is for and in what way it will be used. User characteristics include their profile (such as profession and level of education), which can vary greatly between users, as well as the characteristics of the human visual and cognitive systems, which, although they vary, are more constant.

The context of use is also very important and a visualization may be most suitable for supporting certain users in analyzing a dataset in one circumstance and not suitable in another; for example, the same medical doctor may have to analyze data for a patient sitting in their office or in an intensive care room, and will require different visualization techniques. In recent years, the context of use has become more important in the design of a visualization as it is more frequent to use mobile devices to visualize data, which can occur in a wide variety of situations (such as on the street or in a factory shop floor). This introduces limitations due to the device used (such as screen size, lack of keyboard, processing capacity, or or connectivity issues). Moreover, in the near future, an increase in Immersive Visual Analytics may imply using Extended Reality technologies, introducing other types of limitations and design issues to consider (Skarbez et al.,2019)(Saffo et al. 2023).

This initial design stage is particularly important if the designer/developer is unfamiliar with the application area. The best way to achieve the necessary understanding is, in general, to spend time with potential users to identify their actual needs and how they work, understand the phenomenon under analysis, the data that represents it, the area of application, etc. This work must be complemented by reviewing information on the topic.

1.2 Why: Tasks and Questions

The tasks that the visualizations should support involve specific questions that users will want to answer by analyzing the visualizations (for example, which month or district saw the highest sales of a certain product or how the average seawater temperature has varied over a period of time). Understanding these tasks and questions from the outset allows us to know **why** the visualization is important and what will be used for.

A visualization can effectively be created with many different purposes in view. To communicate, explore, understand, find new information, prove hypotheses, make decisions… A visualization that adequately supports users in a given task may be inadequate in another, for the same dataset. Therefore, user tasks are as important in the design of a solution as the type of data and analyzing them in a more abstract way and independent of the application domain allows considering the similarities and differences in many contexts, so a taxonomy of tasks is a useful approach. This topic is briefly addressed in Section 2.

1.3 What: Data

Before considering representation and presentation, it is necessary to obtain and understand the data to visualize (**what**). These may be "raw" data that were collected from the world, or generated by simulation, treating them and placing them in a suitable format for the use of visualization techniques. If this is not done, not even the most appropriate visualization and interaction techniques will achieve good results.

Several questions should be asked by the developer. Is the phenomenon continuous or discrete? How many attributes does the dataset have? What is the nature of the data to be represented visually? Are they quantities, do they have some intrinsic order (like the days of the week), are they categories (like the names of countries)? Are they georeferenced data (associated with a geographic location), do they express relationships (like connections on a social network), or are they text?

This stage (sometimes referred to as the **mathematics phase** of the visualization process) can be more or less complex and laborious depending on the state of the raw data. It can be quite time-consuming, and may involve the need to create derived attributes, treat a lot of missing data, a lot of outliers, use different units (e.g. of the metric or English systems), and change the format if the current one does not lend itself to the application of visualization techniques. Another question that may arise at this stage is the need to reduce the number of attributes to be visualized (reduction of complexity).

This stage of the visualization process is often underestimated and often considered "less interesting", but it is fundamental and often will consume a more significant amount of time and effort than anticipated. The document "Data-Basic concepts" briefly addresses data processing and includes some bibliographical references to further explore this issue.

1.4 How: Visual Representation

Now, knowing the users, the data available and what tasks the users need to perform, it is necessary to decide how to visually represent the data. As already mentioned, different techniques will be appropriate according to the type of phenomenon and data, the type of user task and context of use. Another important aspect is to consider whether the visualizations will be static or dynamic, i.e., if want users should be able to manipulate the views, for example observing from another angle, or at a different scale, or whether it will even be necessary to allow users to control all the stages of creating a visualization (changing the processing of data, choosing parameters used in the visual representation, for example a color scale of a map or the scale of the axes of a line graph, or even selecting the type of representation to use, for example use an histogram or line graph). Following the InfoVis reference model (Figure 2), this stage can be subdivided into two phases, the **design phase** and the **engineering phase**.

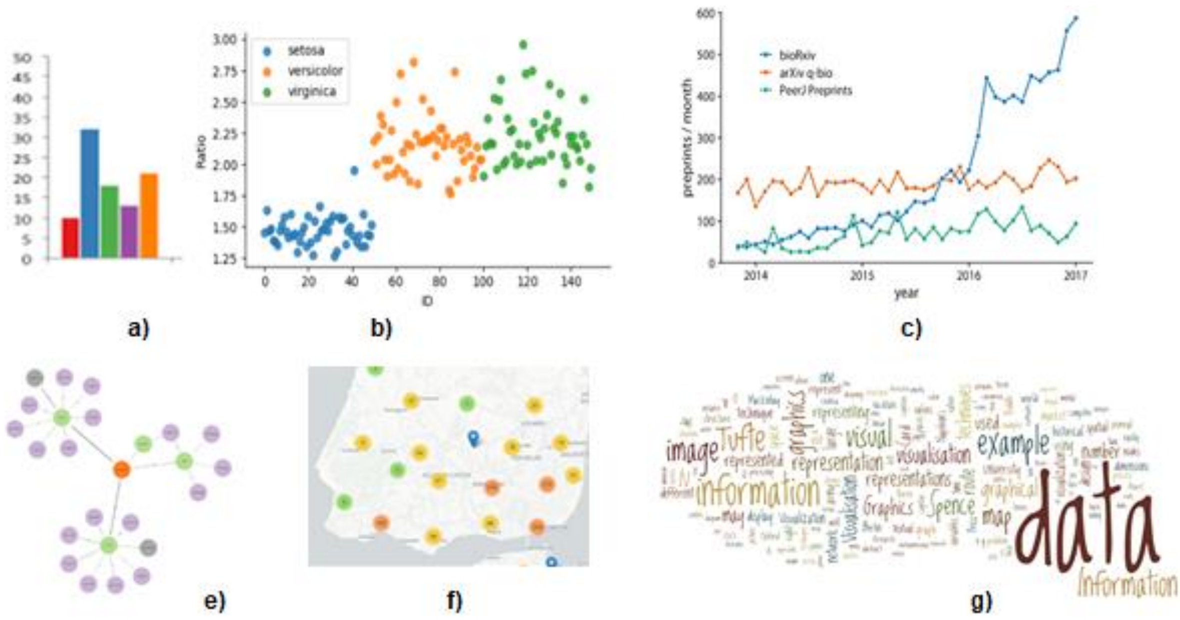

Figure 3. Examples of often used visualization techniques to represent different types of data: a) tabular, b) temporal, c) relation, d) georeferenced, e) text data.

The **design phase** can be considered the core phase of the visualization design process. The selection of the appropriate visualization techniques for each case must take into consideration how people perceive and analyze visual structures in images so that effective and efficient visualizations can be obtained. The topics of data visual encoding, perception and use of color (aspects closely related to the characteristics of the human visual and cognitive system and which establish the necessary foundations for understanding the advantages and limitations of visualization techniques), are presented in detail in (Ware, 2013). The most used visualization techniques to represent the main types of data: tabular and temporal data, relationships, georeferenced data and text, are presented, for instance, in (Munzner,2014) and (Spwncw,2014) illustrating with examples and discussing applicability conditions, advantages and disadvantages. Figure 3 shows examples of visualization techniques often used to represent data of these types.

This phase also includes the design of interaction aspects, closely related to visualization techniques, very relevant as users may need to analyze different facets of the data or using different points of view, especially when static visualizations cannot handle appropriately with the scale and complexity of the data. Taking into consideration the limitations of the human cognitive system (for example the ability to focus attention), trying to show everything simultaneously is not a good solution, even when there are no limits imposed by devices, such as screen size. Issues relating to the selection of interaction techniques are also increasingly important, as it becomes more common to use different types of non-traditional approaches to visualizing data in different contexts (e.g. using virtual and augmented reality or mobile devices). Figure 4 shows some examples of these techniques.

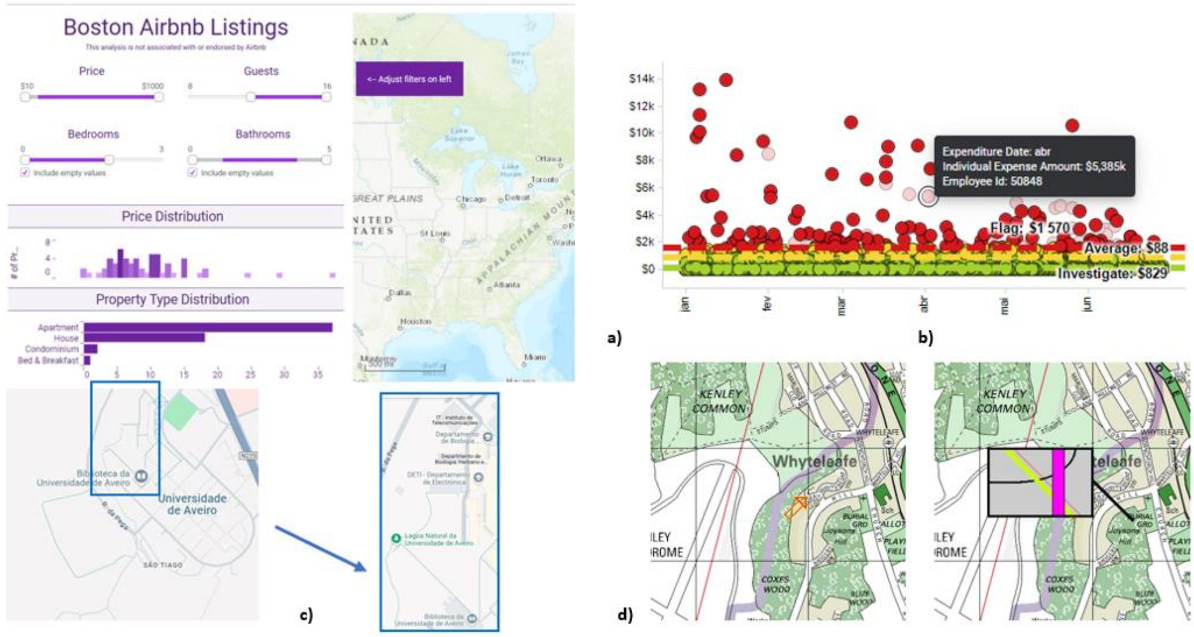

Figure 4. Examples of often used Interaction techniques: a) filtering b) details-on-demand c) semantic zoom d) magic lens[3].

The **engineering phase** can be more or less complex, involving programming skills or not, depending on the software used to produce the visualizations; for example, creating visualizations using Excel or SPSS, does not imply programming, it is only necessary to select in the user interface the correct options that allow to obtain the desired visual representations. However, if we want to develop a visual data exploration application in Google Charts, PowerBI, Matplotlib and D3.js relatively advanced programming skills are needed.

Please note that there is a constant risk in the design of any visualization: the possibility of designing a solution that could mislead users. Although there is no way to guarantee that this will not happen, since users are an integral part of the process and very different from each other, there are principles that, if followed, minimize the possibility of this happening. Design Principles are fundamental and are addressed by several Visualization books (Ware, 2013) (Munzner,2014) (Fisher and Meyer, 2018) (Tufte, 2001), and web sites[4]. An adequate methodology for developing the solution, involving several levels of evaluation and validation,

[3] Adapted from https://www.spotfire.com/demos, google maps, and (Spence, 2007)

[4] https://www.perceptualedge.com/

is also fundamental; this topic is briefly addressed in section 3 and explored in more detail in the document “Evaluation in Visualization”.

In short, a good definition of the problem is fundamental to inform the design of a visualization, and it is good practice to create several alternatives so that those that prove to be best can be chosen through some evaluation process. The design of appropriate solutions will be easier and more likely to be successful, the better the development methodology and the greater the knowledge about existing techniques, their applicability conditions and limitations. However, it will always be necessary to evaluate the solutions and for this there are methods that help selecting the best alternatives for each case.

2. Abstract Task Taxonomy

As already mentioned, visualization can support many different tasks. It can support the discovery (exploration), presentation (explanation), or appreciation (enjoying) of information; it can also support the production of more information for later use. It can be used to generate new hypotheses, such as when exploring a completely unknown dataset, or to confirm existing hypotheses about one that is already partially understood. In other words, a visualization can be used by users to perform data-driven tasks, but a specific visualization can adequately support one task and be completely unsuitable for another, to be performed based on exactly the same dataset, by the same users, in the same context of use. Consequently, it is essential to bear in mind that understanding the tasks that users will perform is as important for the design of a visualization as understanding the type of data in question.

However, when described in terms of the application domain, the tasks can appear very different from case to case, although in general there are many similarities, so a more abstract and application-agnostic description allows to more easily understand which are the most appropriate visualization techniques for each situation. Therefore, it is central to understand how tasks identified in the application domain as being important can be expressed in abstract terms. For example, the tasks of answering the questions “what is the difference between the prices of vegetables sold on local markets and supermarkets?” and “which movies genres were most popular during the year 2020?” may seem very different, but they correspond to comparing values between groups and recognizing that this is a comparison task allows more easily identify which visualization techniques can be used.

Furthermore, considering the tasks that users will perform with the support of the developing visualization in an abstract way and independent of the application domain is also important to better understand how to treat the original data, for example whether it will be useful to derive new attributes. Obtaining derived data is a fundamental process in visualization and can be done through more or less complex arithmetic, logical or statistical operations. Let's look at a simple illustrative case: starting from a dataset of maximum and minimum daytime temperatures recorded throughout the year, in several locations, if one of the tasks that users want to perform is analyzing the variation in thermal amplitude, it may be appropriate to include in the dataset the derived attribute diurnal amplitude (T_{max} -T_{min}).

For these reasons, a taxonomy of tasks is useful. Different taxonomies of visualization tasks have been proposed by several researchers, all describing generic tasks with no connection to the problems or vocabulary of a specific application domain. These tasks describe possible reasons why a visualization should be used. Next the taxonomy proposed by Tamara Munzner (Munzner, 2014) is presented. This taxonomy, in a first level, considers three levels of actions related to users' objectives: analyze, search and consult, indicated in Figure 5.

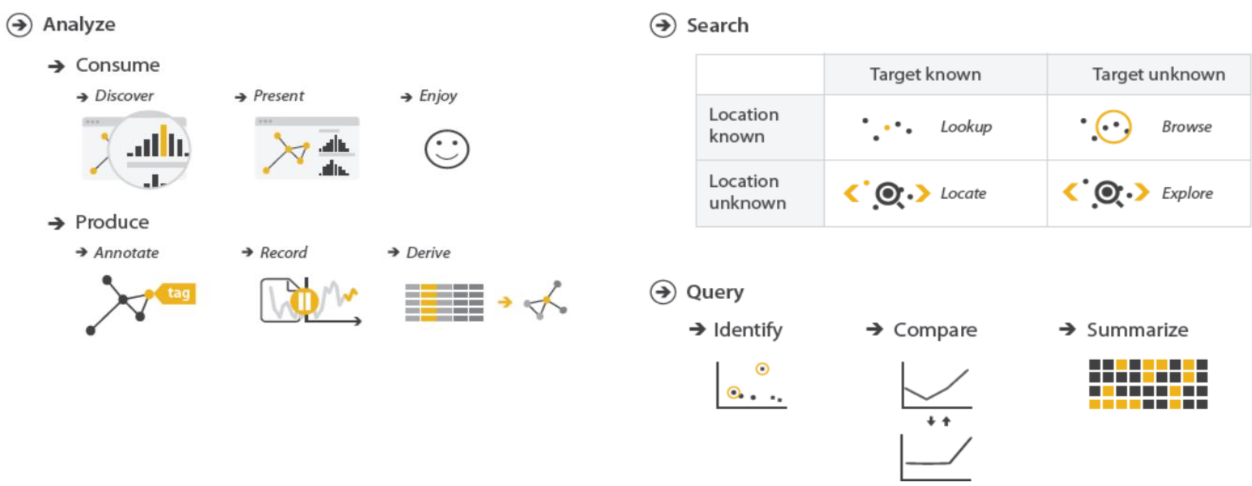

Figure 5. Thee levels of actions (reasons to use a visualization): analyze, search e query (Munzner, 2014)

In the case of Analyze, users may want to consume existing information, such as discovering something new, presenting to others what they already understand, or simply casually enjoying a visualization. For example, and considering again the dataset with maximum and minimum daytime temperatures of various locations, users may want to visually explore the dataset of temperatures recorded in a given location to discover the best time of year to visit that location, to present to others the characteristics of the climate of that location, or without a specific objective, merely to enjoy the view. They may also want to produce new material, temporarily annotating a visualization with graphic elements or text, possibly recording it for later use. They can also produce data derived from the original dataset.

Analyzing a visualization, in general, implies the need to Search for elements of interest. The search task can be divided into four alternatives, Search or Find, if the identity and location of the search object are known, or Navigate and Explore if the search object is unknown.

After elements of interest have been found through search, a frequent task will be to Query these elements, in one of three ways: Identify, Compare or Summarize. While Identify involves only one element (for example finding out which day the maximum temperature was reached during a month), Compare and Summarize refer to two elements or a set of elements, respectively (for example comparing maximum temperatures of two days, or observe the temperatures of every day of a month). Summarizing corresponds to having an overview of the data; it is a very common task in visualization, often being the first approach to visually exploring a dataset; This idea is expressed in the **Information seeking mantra**: “Overview first, zoom and filter, then details on demand” proposed by Ben Shneiderman (Shneiderman,1996).

An abstraction of the tasks provided by this or another taxonomy, together with an abstraction of the data (an issue addressed in the document “Data-Basic Concepts”), are fundamental in the process of designing a visualization application. However, there are “threats to validity” at all stages of the process, meaning it will always be necessary to evaluate and validate the decisions made throughout the process so that the result is useful. In the following section we briefly address this issue.

3. Evaluation and Validation

At each of the stages presented, threats to the validity of the solution under development may occur. In fact, to avoid these threats we must use different approaches to evaluate and validate the decisions made.

At the first stages, there is a risk to not correctly identified the problem to be solved by users using visualization (for example, users will not want to answer the identified questions, but others), which compromises the usefulness of the solution, even if the remaining stages are performed adequately. An appropriate form of evaluation is to use qualitative methods such as interviews with target users or focus groups[5]. A form of further evaluation and validation could be to analyze the adoption rate by users (Munzner, 2014).

At the second and third stages, there is the possibility to have incorrectly identified the types of data or abstract tasks, which will probably result in a choice of inappropriate visualization or interaction techniques in the next stage. One form of evaluation and validation will be testing by users to solve their own problems. To evaluate whether the techniques are appropriate, it is possible to analytical evaluation methods based on known design principles, such as heuristic evaluation[6] or cognitive walkthrough, or carry out studies with users, for example tests based on low-fidelity prototypes or controlled experiments. In the document “Evaluation in Visualization” the most commonly used methods for evaluation and validation in Visualization are presented.

Finally, at the last stage there is the possibility of implementing the techniques using inefficient or ineffective algorithms, resulting, for example, in longer execution time or greater memory consumption than desirable, or lower accuracy in data representation. In this case it is possible to analyze the algorithmic complexity or carry out other types of tests commonly used to validate algorithms in general. To further explore the topic there is bibliography in this area.

[5] https://www.interaction-design.org/literature/article/how-to-conduct-focus-groups

[6] https://www.interaction-design.org/literature/article/heuristic-evaluation-how-to-conduct-a-heuristic-evaluation

Generally, it will be necessary to carry evaluation and validation of the decisions in a later stage, since there are strong interdependencies and decisions taken at one stage will naturally affect the results at subsequent stages. However, in some cases, evaluation and validation can be partially carried out at the same stage; for instance, some evaluation can be performed without implying a full implementation as it is possible to obtain user feedback on selected visualization or interaction techniques based on low-fidelity prototypes. These prototypes can be simply paper prototypes, or can be made using a rapid prototyping tool[7]. Figure 6 shows an example of a low-fidelity prototype for user testing to understand whether selected visualization and interaction techniques support user tasks, before implementing any functional application.

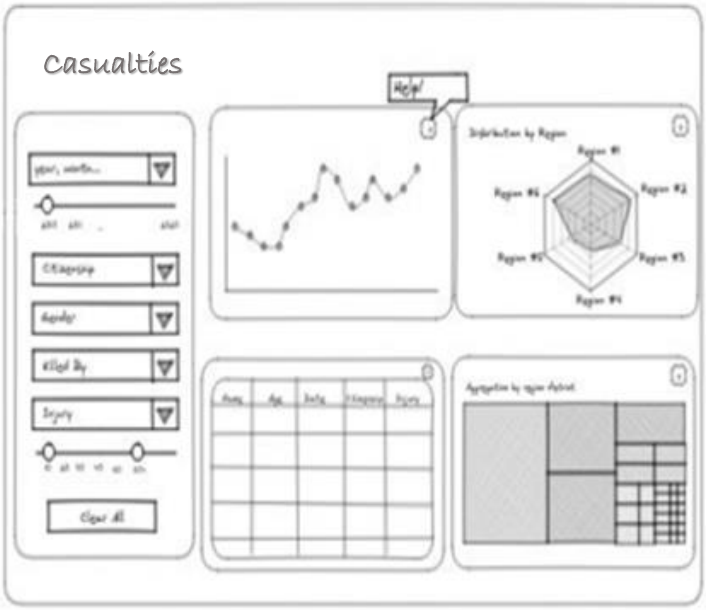

Figure 6. Example of a low fidelity prototype used in usability tests to evaluate the visualization and interaction techniques selected to represent the data and support the user tasks.

References and other Bibliography

- Card, S., Mackinlay, J., Shneiderman, S. (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann.

- Fisher, D. and Meyer, M. (2018). *Making data Visible*. O'Reilly.

[7] https://www.interaction-design.org/literature/article/prototyping-learn-eight-common-methods-and-best-practices

https://learning.oreilly.com/library/view/making-data-visual/9781491960493/cover.html

- Kirk, A. (2019). *Data Visualisation A Handbook for Data Driven Design* (2nd ed.). Sage.
- Kirk, A. (2012). *Data Visualization: A successful design process*. Pack Publishing.

https://learning.oreilly.com/library/view/data-visualization-a/9781849693462/cover.html

- Mazza, R. (2009). *Introduction to Information Visualization*, Springer.
- Munzner, T. (2014). *Visualization Analysis and Design*. A K Peters/CRC Press.

https://learning.oreilly.com/library/view/visualization-analysis-and/9781466508910/

- Saffo, D., Di Bartolomeo, S., Crnovrsanin, T., South, L., Raynor, J., Yildirim, C., & Dunne, C. (2024). Unraveling the Design Space of Immersive Analytics: A Systematic Review. IEEE Transactions on Visualization and Computer Graphics, 30(1), 495–506. https://doi.org/10.1109/TVCG.2023.3327368- Shneiderman, B. (1996). "The eyes have it: a task by data type taxonomy for information visualizations". *Proceedings 1996 IEEE Symposium on Visual Languages*, pp. 336-343.
- Skarbez, R., Polys, N., Ogle, J. T., North, C., & Bowman, D. (2019). "Immersive Analytics: Theory and Research Agenda". *Frontiers in Robotics and AI*, 6 (September)
- Spence, R. (2014). *Information Visualization, An Introduction*. Springer.
- Telea, A. (2015). *Data Visualization, Principles and Practice* (2nd ed.). Routledge.

https://learning.oreilly.com/library/view/data-visualization-2nd/9781466585263/

- Tufte, E. (2001) The Visual Display of Quantitative Information, 2nd ed., Graphics Press, 2001
- Ward, M., G. Grinstein, D. Keim (2010). *Interactive Data Visualization*. A K Peters/CRC Press.
- Ware, C. (2013). *Information Visualization, Perception to Design*, 3nd ed., Morgan Kaufmann.

https://learning.oreilly.com/library/view/information-visualization-3rd/9780123814647/xhtml/CHP003.html