Effect of Reading Habits in Demographics

| Maxwell Júnior  79457  IST – Alameda  maxwell.junior@tecnico.ulisboa.pt | Margarida Morais  86473  IST – Alameda  margarida.p.morais@tecnico.ulisboa.pt | Yasser Zacarias  88647  IST – Alameda  yasser.zacarias@tecnico.ulisboa.pt |
| --- | --- | --- |

# INTRODUCTION

# One of the first things we are taught in school is how to read. And it is this knowledge that allows us to get most of the information in our lives. When learning, most of the time the way we gather the information we need is either through reading books, or in the present times, by reading articles on the internet and searching information through websites (i.e. Wikipedia).

# Not only is reading important when studying, but also, when reading books in our daily lives, and this is because reading helps us develop skills that are important to our wellbeing. It helps to improve your self-expression capabilities since you are extending your vocabulary, and it can teach you how to deal with certain obstacles you find in life or help you to learn a new skill.

# Our desired outcome is understanding how different these habits are between different countries throughout Europe, and also try to find whether are correlations between them or not., if it has an impact in the overall picture of the demographics indicators such as the level of dropout or even high achievement in education system.

# Initially there were many questions about the subject that we wanted to answer through the visualizations, but after further study of the topic, we realized that there was a need to narrow our scope, given our data limitations, and only to focus on the most important questions.

With this in mind, we decided to come out with the questions below, that should be answered by correlating reading habits metrics and some demographic indicators.

**Possible Tasks**

This visualization will allow to:

*Search*

* Search a country in the map and visualize their current information about average time spent reading in minutes.

*Identify*

* Observe a country’s demographic indicators in a certain year.

*Explore*

* Visualize the evolution of the demographic indicators in a specific country through the years.
* Visualize which are the countries with a greater average of income in each education level (basic, intermediate and high).
* Visualize which are the countries with lower and higher percentage of early leavers in education.
* Visualize what is the average of money spent in reading material in a country.

*Compare*

* Compare the demographic indicators and average time spent reading of a country with others.
* Compare the countries with a higher average time spent reading and the countries with more positive demographic indicators.

**Example Questions**

This visualization can help us answer the following example questions:

1. What is the average time spent reading in Spain?
2. What was the average percentage of household expenditure in reading material, in Hungary, in the year of 2016?
3. Given a France’s reading habits, what was the rate of dropout in 2015?
4. What is the average income for a high level of education in Portugal, given the country’s reading habits?

## RELATED WORK

Regarding the inspiration and motivation for this work, it all began with our desire to present a work that could corelate important subjects and perhaps, untapped trends. However, theoretical lectures were also very helpful to gather useful information that guided us to structure the idea of how to encode the data, and what types of idioms to use in order to get the most out of available tools.

We used different sources such as websites, and scientific articles, to gather the highest amount of information, and data to support our subject. This led us to information like the average time spent reading books in some of the countries of European Union available in Eurostat [1] website.

Uma imagem com captura de ecrã

Descrição gerada automaticamente

Figure 1. Example of visualization of time spent reading books in Europe

Regarding implementation, we were able to gather useful information from Eurostat, which in addition to the data, provided some visualizations on it, such as, the visualization of time spent reading books shown in figure 1. This visualization was made using a bar chart, so it did not provide an easy assessment of the difference in terms of minutes, as the interval between minutes was too small, and the colors used had no explicit meaning.

During the laboratories throughout the semester we were able to learn from other groups approach to their subjects and obtain good feedback on how we could improve our work.

**THE DATA**

Our main source of data was the Eurostat website, which is the statistical office of the European Union. Its purpose is to provide high quality statistics regarding Europe, which is why we used so much information from it, in order to be able to have data make comparisons between the EU countries.

**Raw** **Data**

Originally our dataset was composed by 10 different .csv and .xlsx files (approximately 614 KB of crude data), related to multiple countries from a time period of, roughly, 2002 to 2016 – however, the most complete period of time in terms of data was approximately from the 2009 to 2018.

We faced some challenges when acquiring and processing the data since at the beginning we wanted to compare the time spent reding books with other metrics such as the rate of dropout, underachievement of students in reading, mathematics and science, average income by education level. But the lack of variables that would relate these metrics and provide valuable information for the project on the available datasets led us to change the information we wanted to show and focus on the data that was complete.

**Missing Values, and Data Cleaning**

We had multiple data files with heterogenous formats gathered from the same source. One of the initial challenges was to standardize these formats so that the dataset would become coherent.

To ensure the quality of the data needed to answer the raised questions, it was necessary to attend a process of cleaning, that involved eliminating unwanted columns with unnecessary attributes. To solve the issue of having empty cells, we had to decide the best approach to address this problem. For the countries that were in a conflict or got divided along the years, we assigned the value “0” to cells that were empty and when we could not add a value, we erased the country.

Another problem we faced was that most of the data initially had a column named value, which represented for example, a percentage or minutes spent reding. It was necessary to rename each of the value attributes with a name that would identify the domain we were addressing, and not a generic name.

The last part of this standardization process was to alter the name of the countries with their IS02 correspondent code, having already in mind that these codes were what we would use to represent the countries in the visualization graphs.

The process described above was possible by using Pentaho Data Integration as the main tool. Using transformations and other resources available on Pentaho Data Integration. The entire data was then processed and transformed to .json files, for later usage in D3.

## VISUALIZATION

Uma imagem com captura de ecrã

Descrição gerada automaticamente

Figure 2. Overview of the Visualization Layout

**Overall Description**

Our final visualization layout consists in four different idioms, a choropleth map on the top right, a heatmap on the bottom right corner, a stacked bar chart on the top left and a line chart on bottom left corner. The full layout of the visualization is shown above in figure 2.

*Choropleth Map*

Uma imagem com texto, mapa

Descrição gerada automaticamenteThis idiom is one of the most important since it’s the focus of our subject, it represents the data on the average time spent reading in some countries of the European Union.

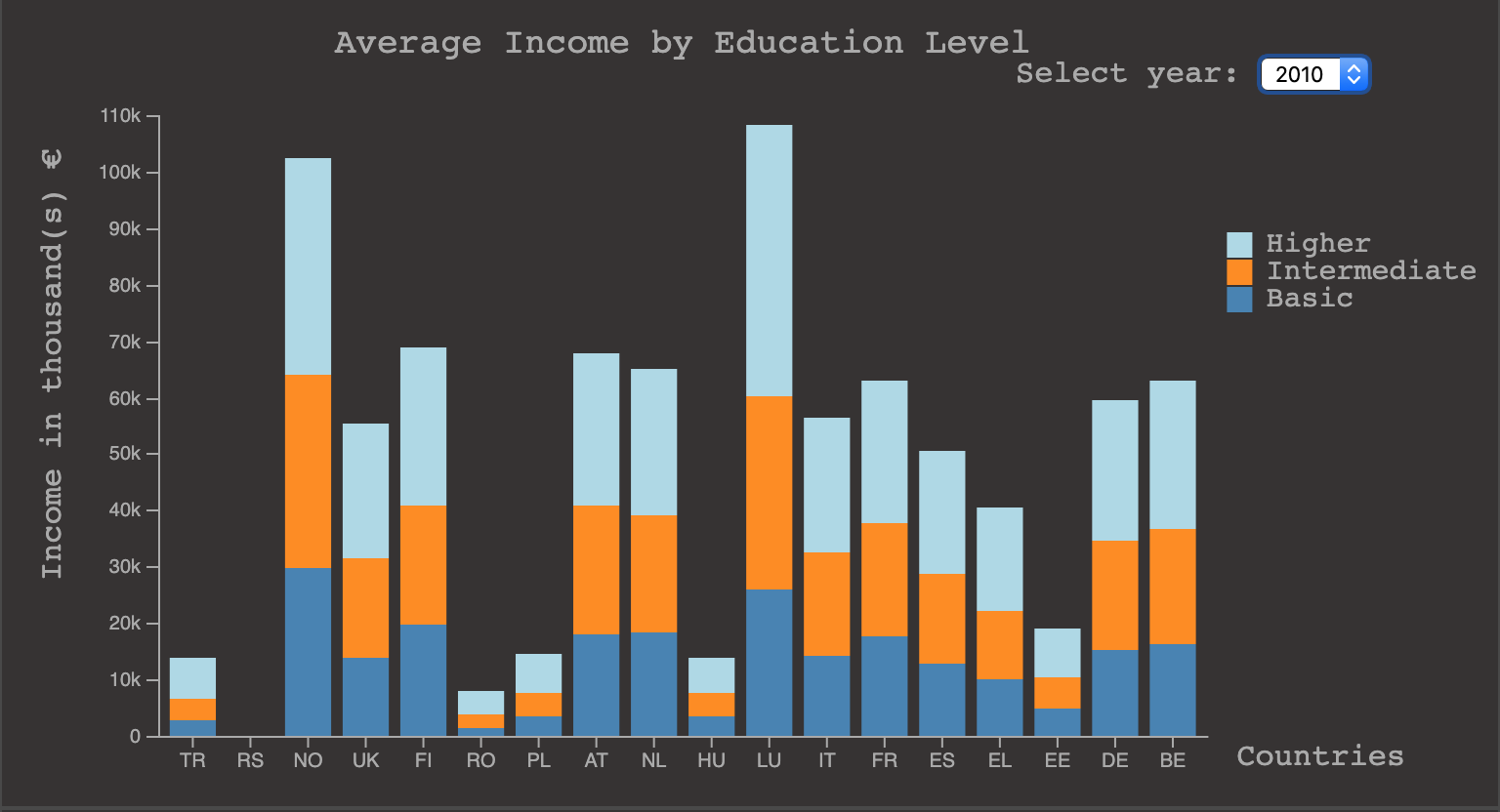
Figure 3. Choropleth Map for Average Time Spent Reading, with mouse pointer hovering over Germany

From the color scale it is possible to understand the differentiation of the countries average time spent reading values, i.e. the strongest colors meaning more time spent and less strong less time. (nao gosto desta descriçao do uso de cor, mas isto também não e saturation, então e o que?)

When the mouse is over a country, it gets a dark color and a pop-up tooltip shows the country’s name and the corresponding average time spent reading in minutes, as seen in figure 3.

*Stacked Bar Chart*

The stacked bar chart allows the user to compare the average income by education level between different EU countries.



**Figure 4. Stacked Bar Chart for Average Income by Education Level**

As it is possible to observe in figure 4, each axis is labeled with the metric it represents, and each bar encodes a country. The three different colors encode the different levels of education.

The usage of the stacked bar chart gives the user the possibility of comparing between the three different levels of education at the same time.

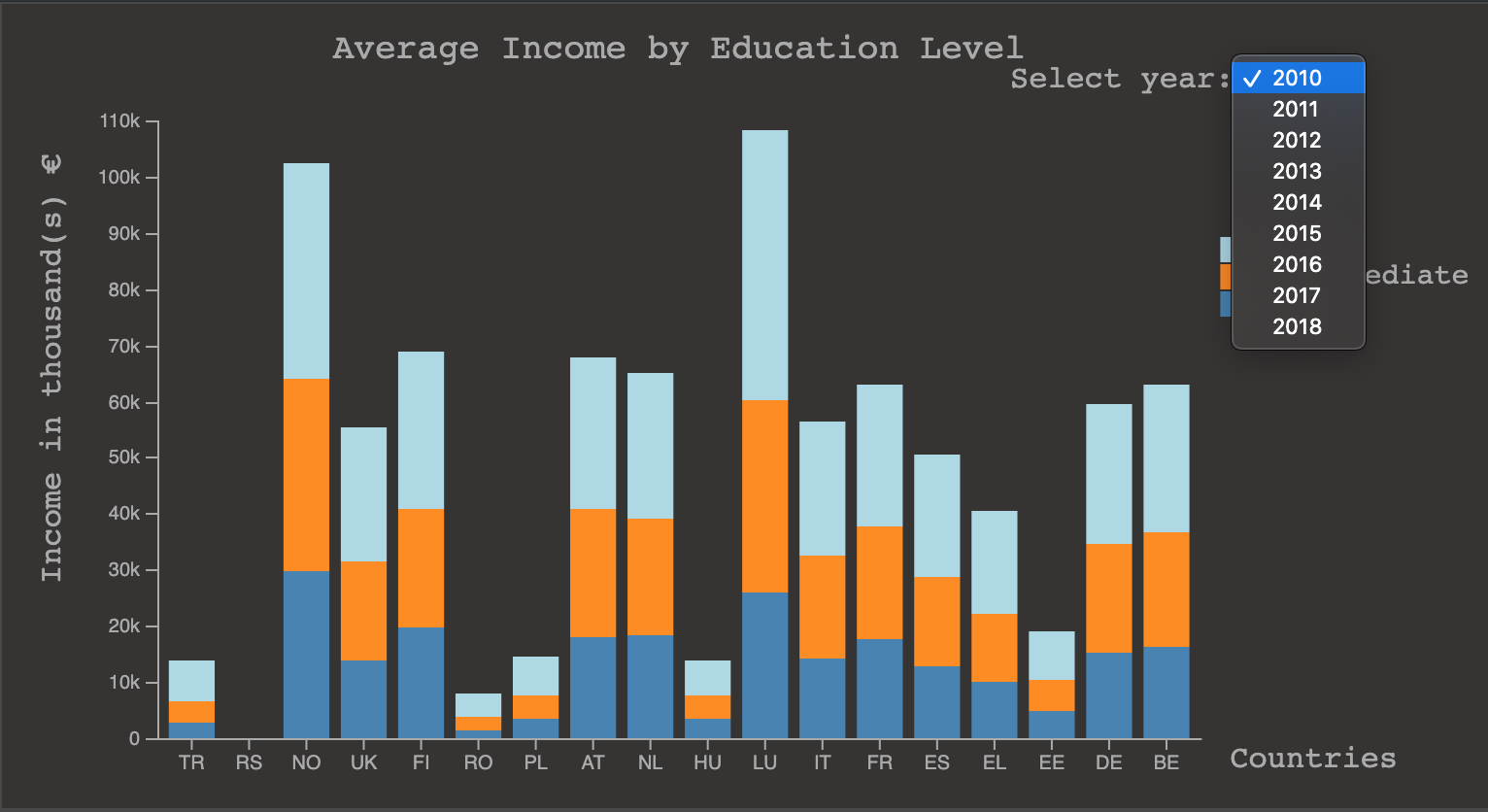


Figure 5. Stacked Bar Chart, filtering through year

The stacked bar char is filtered through the year, the user can select the year s/he wants to observe by selecting it in the dropdown at the right of the graph, just as shown in figure 5.

Uma imagem com captura de ecrã

Descrição gerada automaticamente

Figure 6. Stacked Bar Chart for Average Income by Education Level, with mouse pointer hovering over Finland

When hovering over one of the bars a tooltip will show information about the correspondent country, in this case the value of average income for each education level (figure 6). The tooltip helps understanding the value, since this is hard to perceive simply by looking at the Y axis.

*Line Chart*

The idiom “Education Early Leaver” is represented by a line chart (figure 7) and it allows the user to compare the percentage of early leavers in education between different countries.

Uma imagem com interior, monitor, captura de ecrã, preto

Descrição gerada automaticamente

Figure 7. Line Chart for Education Early Leaver

Each axis is labeled with the metric it represents, and each line represents a country. Each circle encodes the percentage of early leavers in education in that country.

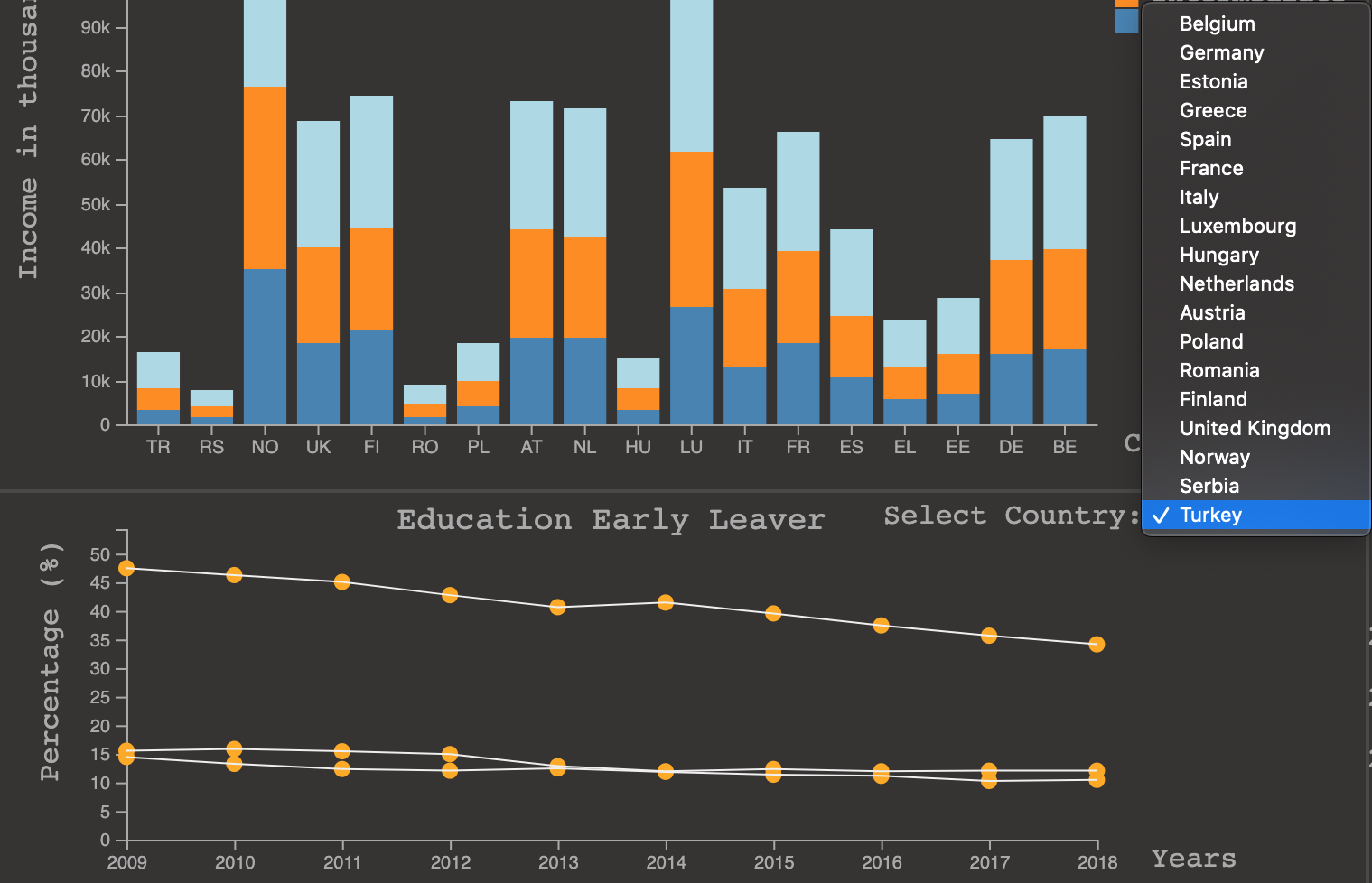


Figure 8. Line Chart, filtering countries to be shown

By selecting a different country in the dropdown list at the right side of the line chart, as shown in figure 8, the selected country correspondent line will be added in the line chart. The user can compare at maximum three countries at the same time.

Uma imagem com monitor, interior, captura de ecrã, ecrã

Descrição gerada automaticamente

Figure 9. Line Chart for Education Early Leaver, with mouse pointer hovering over Belgium in the year of 2016

## By hovering with the mouse over the dots in the line chart, a tool tip will appear showing the country’s name and the correspondent percentage of early leavers in education in that year (figure 9).

*Heatmap*

**Rationale**

For this visualization we used the following encoding:

|  |  |
| --- | --- |
| **Item** | **Visual Encoding** |
| Country | Tag, position (on the map) and color |
| Year | Position and Tag |
| Metrics | Color, tag and slop |

## When we started to work on our layout sketch, one of the first necessary steps was to analyze and select the visualizations that would better suit our data. As a group we had to have several exchanges of ideas and opinions, so that in the end we would come up with the best overall visualization layout.

One of the first visualizations to be chosen was the map, since our subject had the objective of comparing information from different countries.

We chose the choropleth map as it represented a good idiom to overview a metric value for a specific country, as well as comparing it with other countries.

Choosing the stacked bar chart for the visualization of the “Average Income by Education Level” resulted from understanding that we had a lot of variables to encode in the same visualization, and after having a lot of visualization proposals, having the help of the laboratory professor we concluded that using the stacked bar chart would be a good option. Although the stacked bar chart was already representing the metrics country, income in years and education level, we wanted to give the user the possibility of comparison of these metrics between different years, even though we couldn’t have a way of showing information of different years in the same graph, it is possible for the user to select a year and observe the data correspondent to that year.

The percentage of Early Leavers in Education by country was also an idiom where we wanted to provide an easy way for the user to compare information in different countries and be able to visualize the existing evolution through the years. With this idea in mind we started looking at visualizations such as scatter plots, and line charts, in the end, after testing the existing data with different visualizations and prototyping, we concluded that a line chart would be the best option for our goal. One of the first problems that occurred was that, having too many countries represented in the line chart would cause having clutter in the visualization, with this in mind, implementing a filtering tool for the countries solved this problem. Enabling this filtering also provided the necessary interaction with the idiom.

The heatmap representing the average household expenditure in books and newspapers, was one of the last idioms to be decided upon,

Uma imagem com texto

Descrição gerada automaticamente

Initial we come out with different ideas and translate most of them in the first sketch, it was very ambitious, and it seems to be perfect. But when we dive in the project and the implementation technology, we get to understand that was a little bit hard to implement all of the ideas, because it was necessary spend a lot time and work to accomplish every task.

There was only one idiom that ended up being implemented, which is the choropleth map. Because, it was the one that fit in our presentation and for the type of attribute that we were exploiting it works.

For example, at beginning we were trying to implement average income by education level on a heatmap, but after evaluating the quantity of variables that we had to encode, it became clear that the idea did not work.

We also were stuck in the idea of use the parallel coordinates to encode data related with book and newspaper household expenditure, but it was not the best idiom to represent, because after we tried to implement we saw that it could not done with the structure of the lines, crossing each other.

## Why did you think your techniques would work? What visual encodings did you use and why (and why not others)? What alternatives did you consider, even if they turned out not to work? Especially, discuss how you managed the complexity of real data, and matters of scalability. Also, include in your discussion the evolution of the prototype, from the initial sketches to the last version highlighting what you learned from version to version and how that influenced your design.

**Demonstrate the Potential**

## In this section we will demonstrate our visualization implementation and discuss the idioms and the way they interact each other.

To demonstrate the potential of our visualization, we will pick some questions and tasks mentioned in the introduction and show how them can be achieved.

Describe for at least a couple of cases (from the questions you promised you’d answer before) where, step by step (illustrated with screenshots), you find the answers you seek. In short, demonstrate the potential of your solution! Does your visualization provide insights on data that *you were not expecting / that are not common knowledge*? ***These are pure gold!*** Be sure to include them!

**IMPLEMENTATION DETAILS**

**Overall Description**

hat challenges did you find and overcome? How did you implement the links between the views (incl. brushing, etc.)? What algorithms did you use? What techniques did you adapt, or implement, from scratch? (instead of just copying & pasting them from the D3 examples page...)

## References and Citations

Use a numbered list of references at the end of the article, ordered alphabetically by last name of first author, and referenced by numbers in brackets [1,3,4].

Your references should be published materials accessible to the public. Internal technical reports may be cited only if they are easily accessible (i.e., you provide the address for obtaining the report within your citation) and may be obtained by any reader for a nominal fee. Proprietary information may not be cited. Private communications should be acknowledged in the main text, not referenced

References should be in ACM citation format: <http://acm.org/publications/submissions/latex_style>. This includes citations to internet resources [1,4,8,1] according to ACM format, although it is often appropriate to include URLs directly in the text, as above.

# CONCLUSIONS

To conclude, unfortunately we cannot say that this project was successful to attend the expectations that we create around them, having in mind the potential of technologies available.

Although the difficulties that we face to manage our time and the tasks that purpose to do, we enjoy most of the steps of the project, since the concept definition, data collection, to drawing the sketches, and finally hands on implementation. We were also glad to learn a lot of the reading habits of some countries in Europe.

What did you learn? Were you able to address all the questions? If you were to start over, what would you have done differently? Also, if you now had 1 more month and €3000 do spend on this, what else would you do to enrich your solution?

The heading of a section should be in Arial 9-point bold, all in capitals (Heading 1 style). Sections should not be numbered.

## FUTURE WORK

Headings of subsections should be in Arial 9-point bold with initial letters capitalized (Heading 2 style). For sub-sections and sub-subsections, a word like *the* or *of* is not capitalized unless it is the first word of the heading.

### Sub-subsections

Headings for sub-subsections should be in Arial 9-point italic with initial letters capitalized (Heading 3 style).

# Conclusion

It is important that you write for the SIGCHI audience. Please read previous years’ proceedings to understand the writing style and conventions that successful authors have used. State clearly what you have done, not merely what you plan to do, and explain how your work is different from previously published work, i.e., *the unique contribution that your work makes to the field*. Please consider what the reader will learn from your submission, and how they will find your work useful. If you write with these questions in mind, your work is more likely to be successful, both in being accepted into the conference, and in influencing the work of our field.

# REFERENCES

1. Eurostat

https://ec.europa.eu/eurostat