Higher Courses & Unemployment

| Mário Reis  70969  Lisbon, Portugal  mario.fb.reis@gmail.com | Artur Fonseca  75456  Lisbon, Portugal  e-mail address | André Pires  76046  Alenquer, Portugal  e-mail address |
| --- | --- | --- |

# ABSTRACT

The purpose of the project is to learn not only about the unemployment of universities, colleges, courses and courses by area/sub-areas across the years in Portugal, but also the entry level of these public courses in 2016. This report describes the work developed during the first semester in Information Visualization course. It summarizes what we achieved, the problems during the development and future work we could improve in the solution. We will describe every step we took, what we learned from this project and what we pretend to represent.

## Keywords

VIS; Information Visualization; Employment; Unemployment; Higher Education; Public; Private; University; College; Course; d3; Pentaho; Dataset.

# INTRODUCTION

Our domain for the project is the employment/unemployment of higher education courses from all the public and private universities of Portugal during the period of 2007-2015.

We chose to analyze this domain because, as students, it’s a very important topic to us ansd we are concern about which courses/areas have more or less employment/unemployment. We hope to achieve some interesting conclusions about this, clarifying some people and hopefully bring out some curiosity to the rest.

In the beginning we decided to focus on creating a VIS so that we could accomplish 5 tasks on our project:

1. Query->Compare - Compare the unemployment (%) of different courses (regardless of course conclusion year of the graduates).

2. Analyze->Consume->Present – Present the information about unemployment (%) from a specific course graduates across time.

3. Query->Identify – Identify the university with more unemployment (%).

4. Consume->Present – Relation between minimum entry grade and unemployment (%).

5. Query->Summarize – Summarize the employment/unemployment by graduation areas.

Finally, we set 6 representative questions our visualization should give an answer to (for the tasks above):

1. Does Computer Science graduates in IST have more unemployment, in 2015, than Computer Science in ISEL? And in 2007? (Task 1).

2. Is Computer Science in IST having less unemployed graduates in last years? (Task 2).

3. What was the year which had less unemployed people from Computer Science in IST? (Task 2).

4. What is the university with more unemployment? (Task 3).

5. Where the unemployment will be higher? In a course with 14 minimum entry grade or one with 17? (Task 4).

6. What is the graduation area with less/more unemployment? (Task 5).

We did manage to answer all the tasks/questions we proposed ourselves in the first checkpoint; while some of them are answered by only one visualization, others are answered by looking at two graphs, but we made it in an intuitive way in order to improve the user experience.

The main framework we used to develop the graphs for this project was d3 and Pentaho for analyze and treat data in order to come up with interesting results and conclusions for our visualizations.

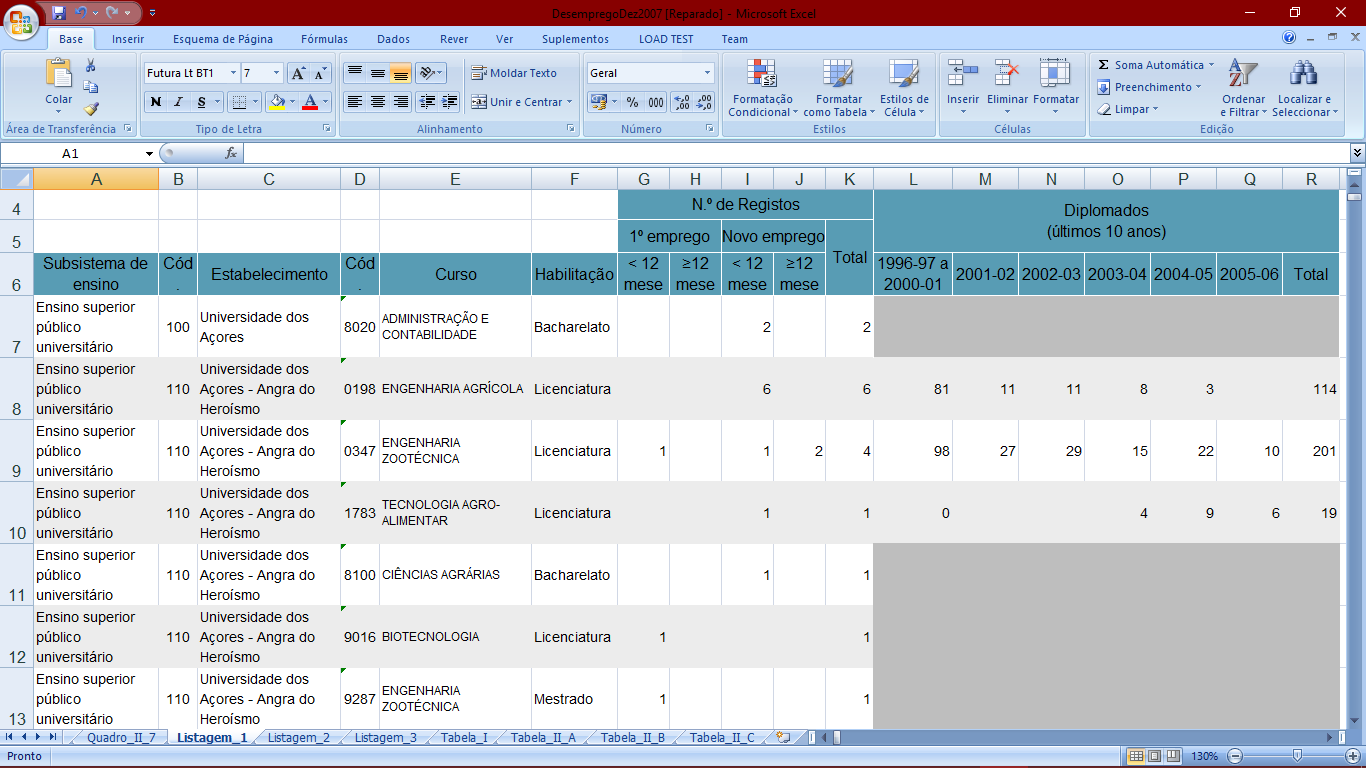
# RELATED WORK

Although we did not find any work like ours, we did find some graphs (mostly scatter plots of line charts) that enlightened us to do our visualizations.

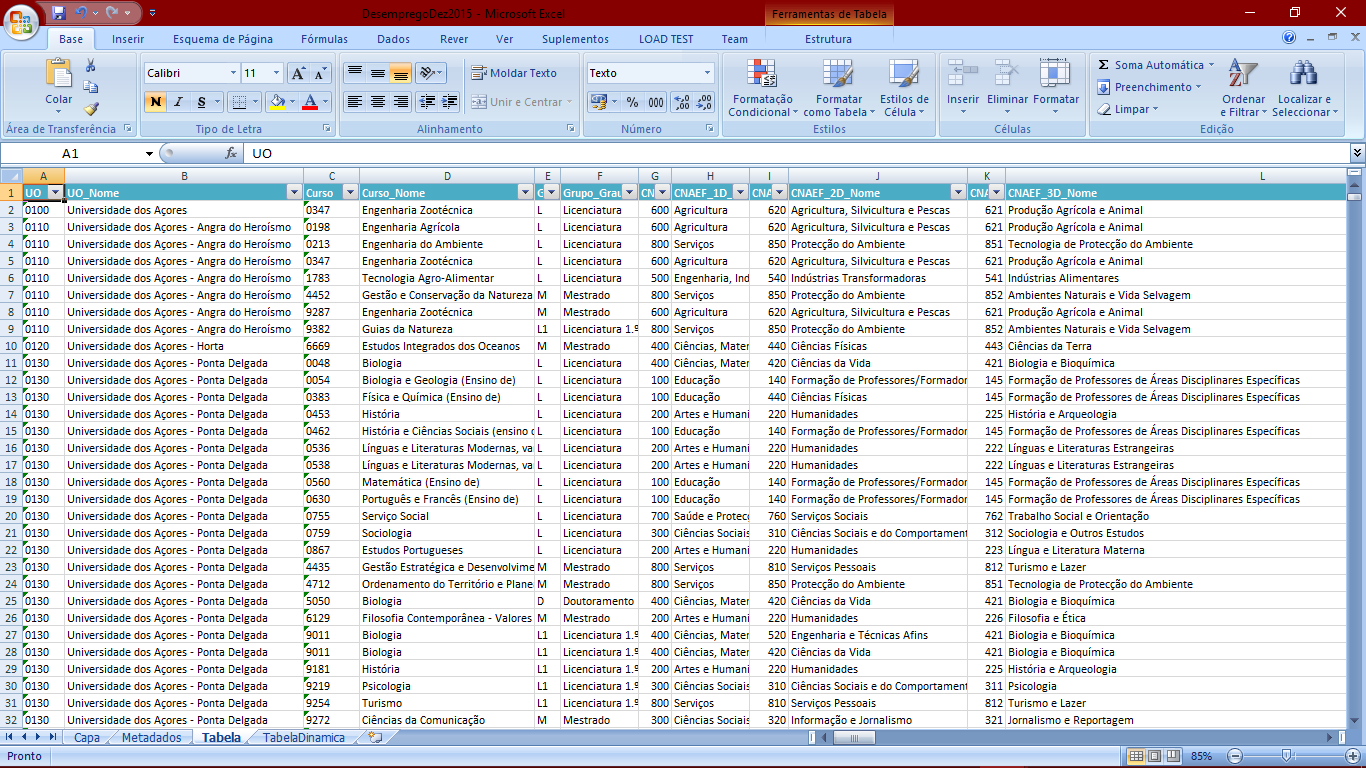
Here are some examples we found interesting and inspiring for our own visualization.

# THE DATA

The data we used was a set of documents (excel format) provided by DGEEC (“Direção-Geral de Estatísticas da Educação e Ciência”) with the information about all the courses and the respective year. This data is available at DGEEC website (<http://www.dgeec.mec.pt/np4/92/>).

Our initial dataset was one file for each year (2007-2015) about unemployment from all the higher education courses registered in “Centro de Desemprego” and other file with entry grades of 2016 for all the higher education courses. The files from 2007–2015 had different layouts and different tables (heterogeneous) as seen in Table 1 and Table 2, which is truncated because it has more than 50 columns/attributes.

**Table 1 – Courses 2007**



**Table 2 – Courses 2015**

We started to select the following attributes: Year, Course Name, Course Code, University Name, University Code, Degree Level, Total Unemployed, Total Graduates, Course Area Code, Course Area Name and Entry Grade.

In particular, there are 3 hierarchical course area levels but top level has many roots “it is a forest of trees”, the figure 1 summarizes the idea (Leaves are the courses).

**Figure 1 – Course Area Levels**

We calculated the following derived measures for each year:

1. % Unemployment by Course (100 \* Total Unemployed of Course/ Total Graduates of Course) to compare the unemployment (%) of different courses (regardless of course conclusion year of the graduates) and to present the information about unemployment (%) from a specific course graduates across time.

2. % Unemployment by Each University (100 \* Total Unemployed of University/ Total Graduates of University) to identify the university with more unemployment (%).

3. % Unemployment By Each Area Level (100 \* Total Unemployed of Area/ Total Graduates of Area) to summarize the employment/unemployment by graduation areas.

Than we decided to spread the information we collected (the attributes referenced above) into three files per year: Courses20XX.json, Areas20XX.json and Universities20XX.json, and also a single file called EntryGrades2016.json, since we concluded that our tasks and questions needed this filtering, which would later on influence the way we present and visualize the information.

We used the table from 2015 to obtain the courses-area relationship, because it was the only one with that information and Merge Join [by Course Code and University Code] it with all the other tables. Some records were lost due to the extinction of courses from “Bolonha” and restructurings (e.g. 2012 dataset had 5100 entries, resulting in 4933 entries).

Some courses didn’t have information about total graduates, we ignored those to make the calculations and aggregations and assigned -1 to the total graduate’s field and unemployment %.

# VISUALIZATION

**Overall Description**

A

**Rationale**

A

**Demonstrate the Potential**

A

# IMPLEMENTATION DETAILS

There were several problems we had to face throughout the semester, but we managed to overcome every one of them.

One of the hardest problems we faced was the modelling of our presentation, what was the best way to show all the graphs, which position they should be and what the best way to depict the data. After discussing we decided we could create an html page where on the left side is Bertin Matrix, on the top-middle is a Scatter Plot, on the top-right is a Sunburst and finally on the bottom is a Line Chart. With only a few dimensions to explore on screens with limited resolution, we had lots of interesting interactions between these idioms to depict the data.

Another big challenge we faced was choosing the right color to represent each course’s area/sub-areas on Sunburst. Each course’s area must have its specific color and each sub-area has its own hierarchical color from its area.

# Conclusion & FUTURE WORK

In the end, we consolidated our knowledge of JavaScript, HTML and CSS. We discovered d3 and learned how powerful this tool is.

We were able to address all the tasks that we set ourselves to do, even though there is always room for improvement.