# Information Visualization

# CHECKPOINT II: Data cleaning and processing

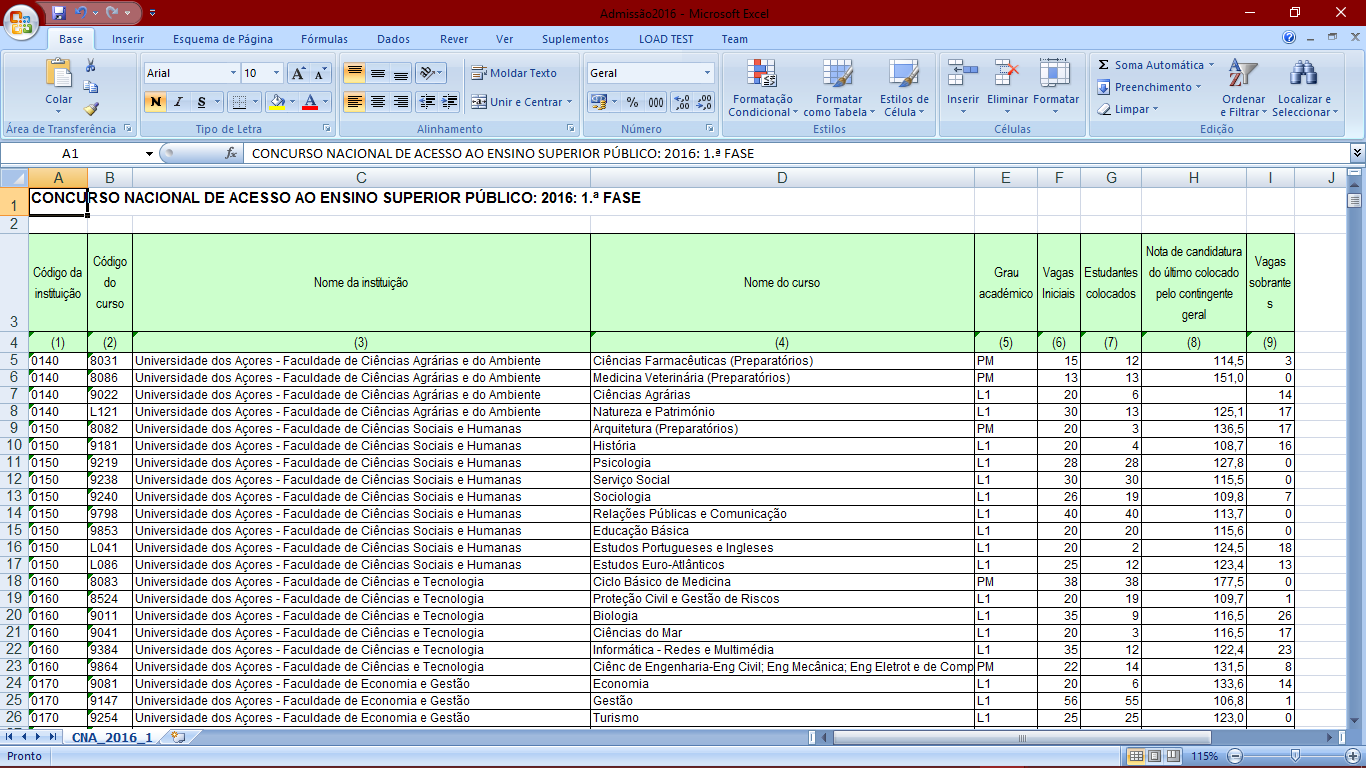
G08 - A

**1. Initial Dataset**

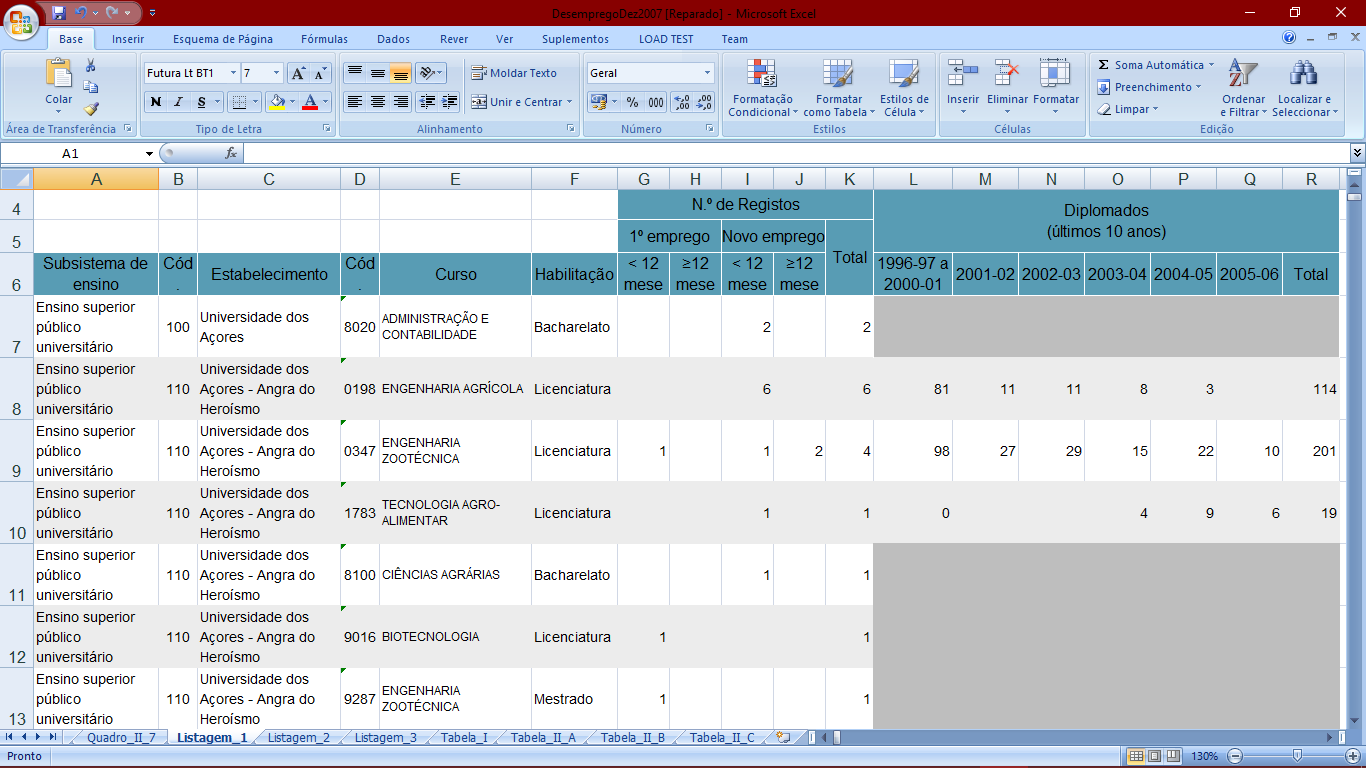
Our initial dataset was one file for each year (2007-2015) about unemployment from all the higher education courses registered in “Centro de Desemprego” and other file with entry grades of 2016 for all the higher education courses.

The files from 2007–2015 each one had different layouts and different tables in short they were a bit heterogeneous as seen in the following samples.

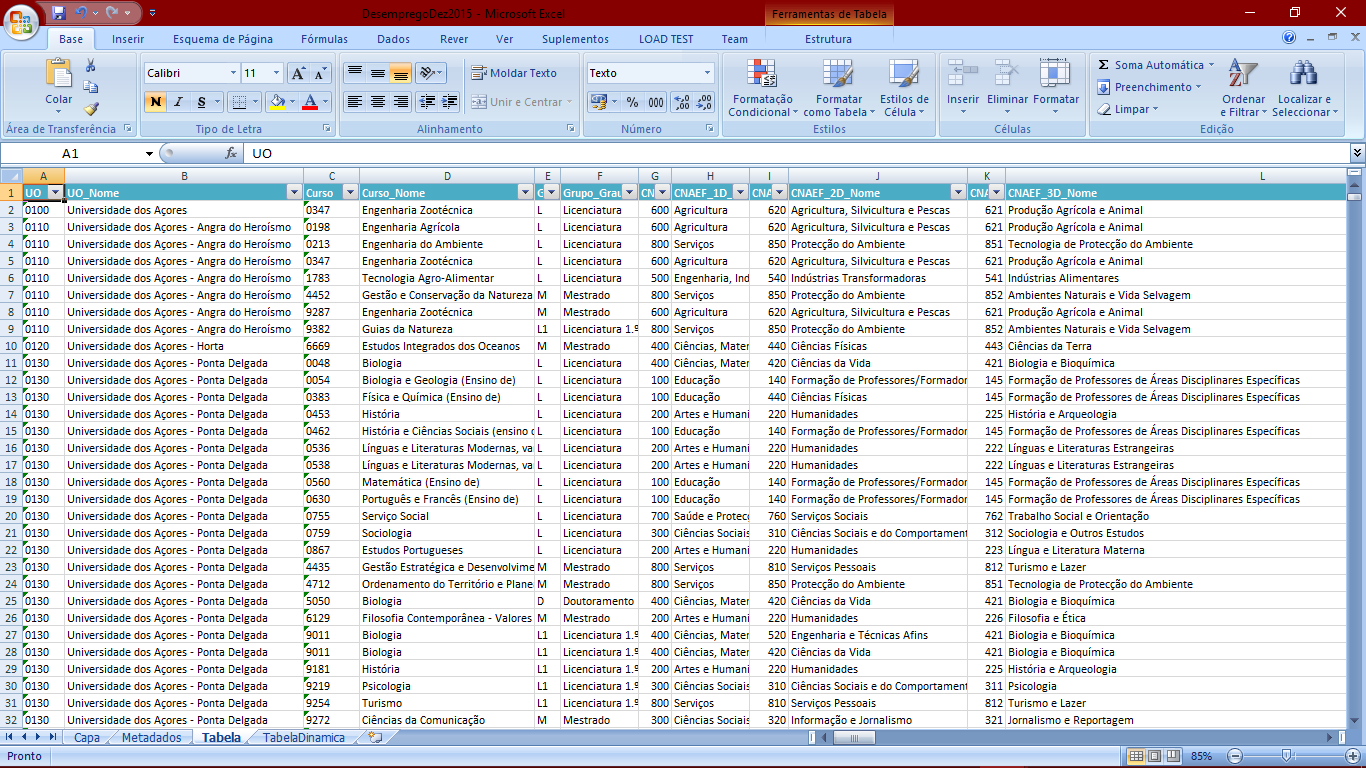
(EntryGrades2016.xls)



(Courses2007.xls)



(Courses2015.xls)



**Note**: This file is truncated because it has more than 50 columns/attributes.

**2. Selected/Derived Data**

We selected the following attributes: **Year**, **Course Name**, **Course Code**, **University Name**, **University Code**, **Degree Level**, **Total Unemployed**, **Total Graduates**, **Course Area Code** and **Course Area Name**

A little explanation about course areas, there are **3 hierarchical levels** but top level has many roots “it is a forest of trees”, the image at right summarize the idea **(Leaves are the courses).**

We calculated the following derived measures **for each year**:

* **% Unemployment By Course** (100 \* Total Unemployed of Course/ Total Graduates of Course) **Task 1**: Compare the unemployment (%) of different courses (regardless of course conclusion year of the graduates) **and** **Task 2**: Present the information about unemployment (%) from a specific course graduates across time
* **% Unemployment by Each University** (100 \* Total Unemployed of University/ Total Graduates of University) **Task 3:** Identify the university with more unemployment (%)
* **% Unemployment By Each Area Level** (100 \* Total Unemployed of Area/ Total Graduates of Area) **Task 5**: Summarize the employment/unemployment by graduation areas

**3. Data abstraction**

Dataset type is **table**. The attributes are the following:

* **Year –** {Continuous|Sequential} It represents the year of the data statistic
* **Course Name/Course Code** – {Nominal} Name of the course/Code of the course
* **University Name/University Code** – {Nominal} Name of the University/Code of the University
* **Degree Level** – {Nominal} If the course is Bachelor’s, Masters...
* **Total Unemployed by Course/Area Level/University** – {Ratio|Sequential} Total number of unemployed of the course/all the unemployed from that area/all the unemployed from a university
* **Total Graduates by Course/Area Level/University** – {Ratio|Sequential} Total number of people that concluded course/all the graduates from courses of the area/all the graduates from courses of the university
* **% Unemployment by Course/Area Level/University** – {Ratio|Sequential} It represents the percentage of unemployed people by course/Area level and University
* **Course Area Name/Course Area Code** – {Nominal|Hierarchical} It represents the code of the course area
  + - 1. **4. Dataset processing**
      2. We used the table from 2015 to obtain the courses-area relationship, because it was the only one with that information and crossed it with all the other tables to obtain all the information we need. We lost some records in 2007 with these due to the extinction of courses from “Bolonha”.
      3. Some courses didn’t have information about total graduates, we ignored these courses to make the calculations and aggregations and assigned -1 to the total graduate’s field and unemployment %.
      4. **5. Mapping (Data sample / Questions)**

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