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 had different layouts and different tables (heterogeneous) as seen in the following samples.

(Courses2007.xls)

						N.º de Registos					Diplomados (últimos 10 anos)						
						1º emprego		Novo emprego		Total							
Subsistema de ensino	Cód.	Estabelecimento	Cód.	Curso	Habilitação	< 12 mese	≥12 mese	< 12 mese	≥12 mese		1996-97 a 2000-01	2001-02	2002-03	2003-04	2004-05	2005-06	Total
Ensino superior público universitário	100	Universidade dos Açores	8020	ADMINISTRAÇÃO E CONTABILIDADE	Bacharelato			2		2							
Ensino superior público universitário	110	Universidade dos Açores - Angra do Heroísmo	0198	ENGENHARIA AGRÍCOLA	Licenciatura			6		6	81	11	11	8	3		114
Ensino superior público universitário	110	Universidade dos Açores - Angra do Heroísmo	0347	ENGENHARIA ZOOTÉCNICA	Licenciatura	1		1	2	4	98	27	29	15	22	10	201

(Courses2015.xls)

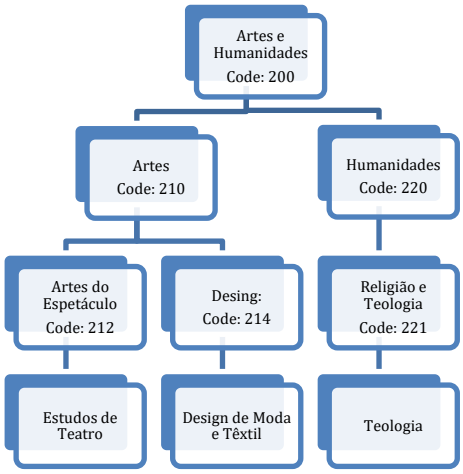
UO	UO_Nome	Curso	Curso_Nome	C	Grupo_Grau	CN	CNAEF_1D	CNA	CNAEF_2D_Nome	CNA	CNAEF_3D_Nome
0100	Universidade dos Açores	0347	Engenharia Zootécnica	L	Licenciatura	600	Agricultura	620	Agricultura, Silvicultura e Pescas	621	Produção Agrícola e Animal
0110	Universidade dos Açores - Angra do Heroísmo	0198	Engenharia Agrícola	L	Licenciatura	600	Agricultura	620	Agricultura, Silvicultura e Pescas	621	Produção Agrícola e Animal
0110	Universidade dos Açores - Angra do Heroísmo	0213	Engenharia do Ambiente	L	Licenciatura	800	Serviços	850	Protecção do Ambiente	851	Tecnologia de Protecção do Ambiente
0110	Universidade dos Açores - Angra do Heroísmo	0347	Engenharia Zootécnica	L	Licenciatura	600	Agricultura	620	Agricultura, Silvicultura e Pescas	621	Produção Agrícola e Animal
0110	Universidade dos Açores - Angra do Heroísmo	1783	Tecnologia Agro-Alimentar	L	Licenciatura	500	Engenharia, Ind	540	Indústrias Transformadoras	541	Indústrias Alimentares
0110	Universidade dos Açores - Angra do Heroísmo	4452	Gestão e Conservação da Natureza	M	Mestrado	800	Serviços	850	Protecção do Ambiente	852	Ambientes Naturais e Vida Selvagem

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**, **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 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|Not Hierarchical} It represents the year of the data statistic

- **Course Name/Course Code** – {Nominal|Not Hierarchical} Name of the course/Code of the course
- **University Name/University Code** – {Nominal|Not Hierarchical} Name of the University/Code of the University
- **Degree Level** – {Ordinal|Not Hierarchical} If the course is Bachelor's, Masters...
- **Entry Grade** – {Ration|Sequencial|Not Hierarchical} Entry grade for the bachelors' course
- **Total Unemployed by Course/Area Level/University** – {Ratio|Sequential|Not Hierarchical} 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|Not Hierarchical} 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|Not Hierarchical} It represents the percentage of unemployed people by course/Area level and University
- **Course Area Name/Course Area Code** – {Nominal|Hierarchical} It represents the name of the course's area/code of the course's area

Important Note: We have this attributes spread in 3 files per year (Courses20XX.json, Areas20XX.json and Universities20XX.json) and a single file called (EntryGrades2016.json).

4. Dataset processing

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 %.

5. Mapping (Data sample / Questions) [new layout]

Does Computer Science graduates in IST have more unemployment, in 2015, than Computer Science in ISEL? And in 2007? **(Task 1)** What was the year which had less unemployed people from Computer Science in IST? **(Task 2)**

Files: Courses20XX.json (attributes: Course Name, University Name, % Unemployment by Course)


```
{"data":[{"NomeFaculdade":"Universidade dos Açores - Angra do Heroísmo","CodigoFaculdade":110,
"CNAEF_3D":621,"NomeCurso":"Engenharia Agrícola","PercentagemDesemprego":6.0,"TotalDiplomados":50,
"TotalDesempregados":3,"Grau":"L","CodigoCurso":"0198"},...
```


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

Files: Universities20XX.json (attributes: University Name, % Unemployment by University)


```
{"data":[{"NomeFaculdade":"DINENSINO-Ensino, Desenvolvimento e Cooperação, CRL (Beja)",
"CodigoFaculdade":20165,"PercentagemDesemprego":2.48,"TotalDiplomados":646,"TotalDesempregados":16},...
```


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

Files: EntryGrades2016.json (attributes: Grade, % Unemployment by Course)


```
{"data":[{"NomeFaculdade":"Universidade do Algarve","CodigoFaculdade":200,"NomeCurso":"Ciências Biomédicas",
"PercentagemDesemprego":7.26,"TotalDiplomados":124,
"Nota":113.5,"TotalDesempregados":9,"Grau":"L1","CodigoCurso":"9351"},...
```


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

Files: Areas20XX.json (attributes: Course Area Name, % Unemployment by Area)


```
{"data":[{"CNAEF":100,"CNAEFNome":"Educação","PercentagemDesemprego":9.05,
"TotalDiplomados":75202,"TotalDesempregados":6803},...
```