Modelação e construção de um *data warehouse*: parte 4

Unidade curricular
Tecnologias de Processamento de Dados - 2017/2018

Docente
António Manuel Silva Ferreira

Grupo 14
Ana Bica, Nº:42333
Henrique Niza, Nº50627
Micaela Domingues, Nº50625
Pedro Ruas, Nº: 50636

Mestrado em Bioinformática e Biologia Computacional

Faculdade de Ciências da Universidade de Lisboa

Índice

1. Fontes de dados

1.1. Dados sobre mobilidade de estudantes Erasmus

Os conjuntos de dados relativos à mobilidade de estudantes no âmbito do programa Erasmus foram fornecidos pelo EU Open Data Portal[1]. O EU Open Data Portal é um repositório de dados abertos sobre políticas de vários organismos dentro da esfera da União Europeia, como instituições, departamentos e agências. Os conjuntos de dados incluem informação do estudante (idade, nacionalidade, instituição de origem, entre outros) e informação acerca da sua participação no programa Erasmus (instituição de acolhimento, país, duração e data da participação, entre outras dimensões). No presente projeto optou-se pelo uso dos conjuntos de dados relativos aos anos de 2009, 2010, 2011 e 2012.

1.2. Dados sobre o *ranking* das universidades

Os conjuntos de dados relacionados com o *ranking* das instituições de ensino superior foram obtidos em SCImago Institutions Rankings[2] (SIR). Os *rankings* são anuais e calculados com base em indicadores de três áreas distintas: desempenho em Investigação e Desenvolvimento (base de dados SCOPUS), inovação (base de dados PATSAT) e visibilidade do trabalho publicado na internet (dados da Google e Ahrefs). Os conjuntos de dados incluem informação sobre a posição que cada instituição ocupa no *ranking* global e o país onde se situa. A fim de se obter uma perspectiva temporal acerca da evolução das universidades europeias no *ranking*, foram usados no presente projecto os conjuntos de dados relativos aos rankings globais para os anos de, 2009, 2010, 2011 e 2012.

1.3. Dados sobre Paridades de Poder de Compra (PPPs)

Relativamente ao conjunto de dados sobre Paridades de Poder de Compra (PPPs), foi usado como fonte de dados a Organisation for Economic Co-operation and Development[3]. O OCDE usa a sua riqueza de informações numa ampla gama de tópicos para ajudar os governos a promover a prosperidade e a combater a pobreza através do crescimento económico e da estabilidade financeira. Ajudam a garantir que as implicações ambientais do desenvolvimento económico e social sejam levadas em consideração. Paridades de poder de compra (PPPs) são as taxas de conversão de

moeda que igualam o poder de compra de moedas diferentes, eliminando as diferenças nos níveis de preços entre países. Na sua forma mais simples, as PPPs mostram a proporção de preços em moedas nacionais do mesmo bem ou serviço em diferentes países. Este indicador é medido em moeda nacional por dólar. O conjunto de dados inclui informações sobre o país, ano (2009 a 2012) e valor do indicador PPP.

1.4. Dados sobre Populações Nacionais

O repositório World Bank's Open Data4 providencia acesso à base de dados "World Development Indicators". Esta base de dados contém um conjunto de indicadores acerca do desenvolvimento a nível global, nacional e regional. Para o projecto, foram extraídos dados relativos à população total dos países para os anos de 2009, 2010, 2011 e 2012.

1.5. Dicionário Erasmus ID codes

Os códigos de dados relativos à mobilidade de estudantes no âmbito do programa Erasmus foram fornecidos pela European Commission Education, Audiovisual and Culture Executive Agency[5]. Os dados de cada participação de Erasmus, relativamente à cidade, país instituição, entre outros estão codificados em código de três ou quatro letras.

1.6. Dicionário ISO 3166 Codes

O ISO 3166 é o padrão internacional para códigos de países e suas subdivisões. Foram obtidos no RIPE Network Coordination Centre[6]. O objetivo do ISO 3166 é definir códigos de letras e/ou números reconhecidos internacionalmente que podem ser usados ao referirmos aos países e subdivisões. O uso de códigos facilita e evita erros ao contrário de usar o nome do país uma vez que este muda dependendo da linguagem usada e não é entendido em todo o mundo.

1.7. Dicionário Erasmus Subject Area Codes (ISCED97 codes)

Os códigos Erasmus Subject Area foram obtidos na ERASMUS+ - Nicolaus Copernicus University - UMK[7]. Os códigos ISCED97 funcionam como uma referência internacional de classificação das áreas de estudo ou estágio

2. Lista de campos

2.1 Dados sobre mobilidade de estudantes Erasmus

Tabela 1: Descrição dos dados sobre mobilidade de estudantes.

Campo	Tipo de Dados	Descrição	Exemplo
HOMEINSTITUTION	Texto	Código de identificação Erasmus da Instituição de origem do aluno. O seu domínio consiste numa sequência de até 13 caracteres no conjunto de valores possíveis denominados "Códigos de identidade Erasmus".	P LISBOA04
COUNTRYCODEOFHOMEINSTITUTION	Texto	O código do país da instituição de origem onde o estudante estuda/está registrado. O seu domínio consiste numa sequência de 2 caracteres no conjunto de “ISO Country Codes”.	PT
AGE	Número inteiro	A idade do aluno definida como sendo a diferença entre o ano do início do Ano Erasmus, e o ano em que o aluno nasceu. O seu domínio consiste num integer entre 16 e 99.	21
GENDER	Texto	Género do aluno. O seu domínio consiste num caractere com letras maiúsculas (M, F). M=Masculino, F=Feminino.	F
NATIONALITY	Texto	Nacionalidade do aluno. O seu domínio consiste numa sequência de 2 caracteres no conjunto de “ISO Country Codes”.	FR
SUBJECTAREA	Número inteiro	A área de estudo do estudante na HOMEINSTITUTION. O seu domínio consiste num integer com máximo de 4 caracteres de acordo com o código designado ISCED97 codes ou Erasmus subject area codes.	34
LEVELSTUDY	Texto	Grau dos estudos na instituição de origem (Ciclo). O seu domínio consiste num caractere com 4 valores possíveis: 1=Primeiro Ciclo, 2=Segundo Ciclo, 3=Terceiro Ciclo, S=Ciclo curto.	1
YEARSPRIOR	Número inteiro	Número de anos completados de educação superior previamente ao período no estrangeiro. O seu domínio consiste num integer entre 0 e 20.	2
HOSTINSTITUTION	Texto	A instituição na qual o estudante passou o período de Erasmus. O seu domínio consiste numa sequência de até 13 caracteres no conjunto de valores possíveis denominados "Códigos de identidade Erasmus".	F POITIER01
COUNTRYCODEOFHOSTINSTITUTION	Texto	O código do país da instituição onde o estudante passou o período de Erasmus. O seu domínio consiste numa sequência de 2 caracteres no conjunto de “ISO Country Codes”.	DK
LENGTHSTUDYPERIOD	Número decimal	Duração do período Erasmus em meses. O seu domínio consiste num valor numérico, com duas casas decimais, entre 0.00 e 12.00. A unidade de medida mais pequena é 0.25, pelo que valores tais como 3.40 não são válidos.	3.5
STUDYSTARTDATE	Data	Data de início do período Erasmus, com formato mm/yyyy.	set/09
MONTH	Texto	Mês da data de início do período Erasmus.	set
YEAR	Número inteiro	Ano da data de início do período Erasmus.	9
TOTALECTSCREDITS	Número inteiro	Número antecipado de ECTs feitos pelo estudante durante o período Erasmus. O seu domínio consiste num inteiro entre 0 e 90.	30
SEVSUPPLEMENT	Número decimal	Bolsa concedida por motivos de deficiência. O seu domínio consiste num valor positivo da moeda (euro), com até duas casas decimais.	3124.22

TAUGHTHOSTLANG	Texto	Reporta se a aprendizagem foi feita na língua do país anfitrião. O domínio consiste num caractere em maiúsculas (Y, N), Y=sim, N=não.	Y
LANGUAGETAUGHT	Texto	A linguagem na qual se efetuou a aprendizagem. O seu domínio consiste numa sequência de dois caracteres, de acordo com o conjunto “ISO Language Codes”.	EN
LINGPREPARATION	Texto	Reporta se o estudante frequentou um curso de línguas no país anfitrião ou de origem, ou um outro tipo de curso relacionado com Erasmus. O seu domínio consiste uma sequência de dois caracteres no conjunto (EC, HS, HM, NN), em que EC = EILC, HS = Anfitrião, HM = Origem, NN = Nenhum.	NN
STUDYGRANT	Número decimal	Bolsa recebida pelo estudante, excluindo a bolsa de deficiência. O seu domínio consiste num valor positivo da moeda (euro), com até duas casas decimais.	630.00
PREVIOUSPARTICIPATION	Texto	Reporta participação prévia no programa Erasmus. O seu domínio consiste num caractere de letras maiúsculas (N,S,P,M), sendo que N=não, S=estudo, P=estágio, M=Erasmus Mundus.	N
QUALIFICATIONATHOST	Texto	Reporta a qualificação que o estudante receberá na instituição anfitriã. O seu domínio consiste num caractere de letras maiúsculas (D, J, O, N), em que D=diploma duplo, J=diploma em conjunto, O=Outro, N=Nenhum (Europass, etc.).	N

2.2 Dados sobre o *ranking* das universidades

Tabela 2: Descrição dos dados sobre *raking* de instituições.

Campo	Tipo de Dados	Descrição	Exemplo
Global.Rank	Número Inteiro	Posição no *ranking* mundial que a instituição ocupa. O seu domínio consiste num inteiro que pode tomar valores de 1 até um valor variável, que corresponde ao número de universidades classificadas em cada ano.	1
Institution	Texto	Nome da instituição que se encontra no ranking. O seu domínio é uma *string* de dimensão variável.	Centre National de la Recherche Scientifique
Country	Texto	Código que identifica o país a que a instituição pertence. O código é constituído por uma *string* de três letras correspondente ao respetivo código A3 do Dicionário ISO 3166 Codes para um dado país. O seu domínio restringe-se aos países participantes no programa Erasmus: AUT, BEL, BGR, CYP, CZE, DEU, DNK, EST, ESP, FIN, FRA, GBR, GRC, HRV, HUN, IRL, ISL, ITA, LIE, LTU, LUX, LVA, NLD, NOR, POL, PRT, ROU, SWE, SVN, SVK, TUR.	DEU
Sector	Texto	Sector a que pertence a instituição. O seu domínio é uma string que pode adquirir os valores: “Government”, “Health”, “Higher Educ.”, “Others”, “Private”.	Government
Year	Data	Ano a que se refere a classificação no ranking. O seu domínio é: 2009 - 2012.	2009

2.3 Dados sobre Paridades de poder de compra (PPPs)

Tabela 3: Descrição dos dados sobre paridades de poder de compra.

Campo	Tipo de Dados	Descrição	Exemplo
COUNTRY	Texto	Código único de 3 letras para identificar o nome de um país (correspondente ao respetivo código A3 do Dicionário ISO 3166 Codes para um dado país)	AUS
YEAR	Data	Ano correspondente ao indicador em formato YYYY.	2010
VALUE	Número decimal	Unidades monetárias nacionais por dólar dos EUA, valor PPP. O seu domínio consiste num valor positivo até seis casas decimais.	0.862944

2.4 Dados Populações Nacionais

Tabela 4: Descrição dos dados sobre populações nacionais.

Campo	Tipo de Dados	Descrição	Exemplo
Country Name	Texto	Nome do país (sequência de caracteres).	Belgium
Country Code	Texto	Código do país, que consiste numa sequência de 3 caracteres no conjunto de “ISO Country Codes”.	BEL
Year	Data	Ano em que foi registado o indicador, no formato yyyy.	2011
Population	Número inteiro	Tamanho populacional.	10796493

2.5 Dicionário “Erasmus ID codes”

Tabela 5: Descrição dos dados do dicionário de códigos Erasmus.

Campo	Tipo de Dados	Descrição	Exemplo
Country	Texto	País da instituição O seu domínio consiste numa sequência de 2 caracteres no conjunto de “ISO Country Codes”.	IT
Charter type code	Texto	Medida que fornece o quadro geral de qualidade para as atividades de cooperação europeias e internacionais que uma instituição de ensino superior pode realizar no âmbito do Erasmus. O domínio consiste em numa sequência de 3 caracteres no conjunto (EUC, EUCX, EUCP); EUC=standard, EUCX=extended e EUCP=placement.	EUCX
Organization Name	Texto	Nome da instituição (sequência de caracteres).	UNIVERSITA' DEGLI STUDI DI NAPOLI FEDERICO II
Erasmus code	Texto	Código da instituição, segundo o conjunto de “Erasmus ID codes”.	I NAPOLI01
Street	Texto	Morada da instituição (sequência de caracteres).	Corso Umberto I - 40 bis
Postcode	Texto	Código postal da instituição	4382 NW
City	Texto	Nome da cidade da instituição (sequência de caracteres).	Napoli

2.6 Dicionário ISO 3166 Codes:

Tabela 6: Descrição dos dados do dicionário de códigos de países.

Campo	Tipo de Dados	Descrição	Exemplo
Country	Texto	Nome do país (sequência de caracteres < 50).	DENMARK
A2	Texto	Código único de duas letras para identificar o nome de um país.	DK
A3	Data	Código único de 3 letras para identificar o nome de um país.	DNK
Number	Número inteiro	Código único de 3 algarismos para identificar um território físico.	208

2.7 Dicionário Erasmus Subject Area Codes (ISCED97 codes)

Tabela 7: Descrição dos dados do dicionário Erasmus com as áreas de estudo.

Campo	Tipo de Dados	Descrição	Exemplo
Equivalent ISCED97	Número inteiro	Número inteiro único para identificar a área.	462
ISCED97 Description	Texto	Descrição da área em questão.	Statistics

3. Preparação e limpeza de dados

De modo a garantir a qualidade dos dados extraídos das fontes referidas, procedeu-se inicialmente à deteção de erros nos conjuntos de dados selecionados, seguida das tarefas de preparação e limpeza. Estas tarefas são essenciais para a concretização dos objetivos propostos no âmbito do projeto.

3.1 Dados sobre mobilidade de estudantes Erasmus

Os dados relativos à mobilidade de estudantes Erasmus são referentes, como anteriormente descrito, aos anos 2009, 2010, 2011 e 2012. Os dados estão agrupados em ficheiros individuais para cada ano letivo: 2009-2010, 2010-2011 e 2011-2012. A uniformização dos dados foi conseguida através das seguintes transformações:

- Eliminação de colunas: MOBILITYTYPE, WORKPLACEMENT, ENTERPRIZESIZE, TYPEWORKSECTOR, LENGHTWORKPLACEMENT, SHORTDURATION, PLACEMENTSTARTDATE e CONSORTIUMAGREEMENTNUMBER por omissão completa ou parcial de valores. A dimensão ECTSCREDITSWORK foi eliminada, uma vez que seria despropositada sem as anteriores dimensões eliminadas. A dimensão ECTSCREDITSSTUDY foi eliminada por ser redundante relativamente à TOTALECTSCREDITS;
- Eliminação de linhas com campos não preenchidos (em branco);
- Eliminação de linhas com campos preenchidos com ‘XX’;
- Uniformização do código de país para Bélgica. Observou-se que, ao longo dos dados o mesmo país era referido através de 3 códigos diferentes, BE, BEFR e BENL, pelo que se passou a utilizar apenas um código, BE;
- Eliminação de linhas cuja duração do intercâmbio tinha o valor 0;
- Uniformização de casas decimais nos campos com valores numéricos;
- Uniformização das datas para um formato reconhecível pela base de dados, por exemplo, de “jan/2010” para “jan-2010”.
- Eliminação/substituição de caracteres fora do domínio utf-8, nomeadamente, no campo Subject Area Name do dicionário de Subject Area, e no dicionário Erasmus code, nos campos Organisation name, street e postcode.
-

3.2 Dados sobre o *ranking* das universidades

Como já foi referido anteriormente, a informação sobre os *rankings* das universidades para cada ano encontrava-se dispersa por 4 diferentes conjuntos de dados, correspondentes ao *ranking* global para os anos de 2009, 2010, 2011 e 2012.

Para manipular os dados de forma sistemática e simplifica foi desenvolvido um script em Rstudio (ver ficheiros em anexo: “rankingNotebook.nb.html” e “rankingNotebook.rmd”). De um modo geral, o *script* referido importa dados de vários ficheiros .csv, compila os dados todos numa única variável, introduz um novo campo “Year”, relativo ao ano a que se refere a posição no ranking global, selecciona do *ranking* global apenas as linhas correspondentes a instituições de países participantes no programa Erasmus, elimina a primeira linha, que se encontra em branco e produz um ficheiro output “RankingAnos.csv” com todos os dados.

3.3 Dados sobre Paridades de Poder de Compra (PPPs)

Foram eliminadas colunas de dados redundantes, nomeadamente INDICATOR, SUBJECT, MEASURE, FREQUENCY e a coluna FLAG por conter células vazias.

4. Análise de estatística descritiva

4.1 Dados sobre Erasmus

Tabela 8: Análise de estatística descritiva de dados sobre mobilidade de estudantes. Para dados categóricos foram obtidos os valores mais importantes, juntamente com uma contagem do número de vezes que cada valor ocorre; para dados contínuos, foi feita a caraterização da gama de valores, incluindo o mínimo, mediana, média e máximo.

Campo	Tipo de Dados	Análise Estatística
HOMEINSTITUTION	Categóricos	E GRANADA01: 5697
		E MADRID03: 5148
		I BOLOGNA01: 4492
		E SEVILLA01: 4190
		E VALENCI01: 3614
		I ROMA01: 3527
		(Other): 534027
COUNTRYOFHOMEINSTITUTION	Categóricos	ES: 91908
		DE: 75172
		FR: 74660
		IT: 57902
		PL: 34806
		TR: 27069
		(Other): 199178
AGE	Contínuos	Min. :17.00
		1st Qu.:21.00
		Median :22.00
		Mean :22.49
		3rd Qu.:23.00
		Max. :99.00
GENDER	Categóricos	F:340715
		M:219980
NATIONALITY	Categóricos	ES: 91156
		DE: 79099
		FR: 74927
		IT: 58935
		PL: 36157
		TR: 27780
		(Other): 192641
SUBJECTAREA	Categóricos	222 (Foreign languages) : 82275
		22 (Humanities): 58388
		34 (Business and administration): 41173
		340 (Business and administration (broad programmes)): 37259
		314 (Economics): 23988
		313 (Political science and civics): 20055
		(Other): 394802
LEVELSTUDY	Categóricos	1:391635
		2:161365
		3:4817
		S: 2878
YEARSPRIOR	Contínuos	Min. : 1.000
		1st Qu.: 2.000
		Median : 2.000
		Mean : 2.756
		3rd Qu.: 3.000

		Max. :20.000
HOSTINSTITUTION	Categóricos	E GRANADA01: 5871
		E VALENCI01: 5104
		E MADRID03: 4898
		E SEVILLA01: 4738
		I BOLOGNA01: 4734
		E VALENCI02: 4585
		(Other): 530765
COUNTRYOFHOSTINSTITUTION	Categóricos	ES: 89892
		FR: 67597
		DE: 56945
		UK: 51501
		IT: 49356
		SE: 26444
		(Other): 218960
LENGTHSTUDYPERIOD	Contínuos	Min. : 0.250
		1st Qu.: 4.500
		Median : 5.000
		Mean : 6.361
		3rd Qu.: 9.000
		Max. :13.250
STUDYSTARTDATE	Discretos	Sep-11 : 100795
		Set-10 : 92078
		Set-09 : 86287
		Feb-12 : 27054
		Fev-10 : 25032
		Fev-11 : 24485
		(Other): 204964
MONTH	Discretos	set: 279277
		jan : 54471
		fev : 49532
		out : 39322
		ago : 38736
		mar: 16417
		(Other): 99357
YEAR	Discretos	9: 125563
		10: 182950
		11: 195702
		12: 56478
TOTALECTSCREDITS	Contínuos	Min. : 0.00
		1st Qu.:24.00
		Median :30.00
		Mean :34.48
		3rd Qu.:48.00
		Max. :90.00
SEVSUPPLEMENT	Contínuos	Min. : 0.00
		1st Qu.: 0.00
		Median : 0.00
		Mean : 2.93
		3rd Qu.: 0.00
		Max. :41777.74
TAUGHTHOSTLANG	Categóricos	N:222088
		Y:338607
LANGUAGETAUGHT	Categóricos	EN: 276538
		ES: 77103

		FR: 67080
		DE: 52518
		IT: 41507
		PT: 13035
		(Other): 32914
LINGPREPARATION	Categóricos	EC: 15184
		HM:104926
		HS: 92879
		NN:347706
STUDYGRANT	Contínuos	Min. : 0
		1st Qu.: 925
		Median : 1259
		Mean : 1483
		3rd Qu.: 1844
		Max. :10970
PREVIOUSPARTICIPATION	Categóricos	M: 249
		N:558006
		P: 1861
		S: 579
QUALIFICATIONATHOST	Categóricos	D: 7122
		J: 3089
		N:532550
		O: 17934

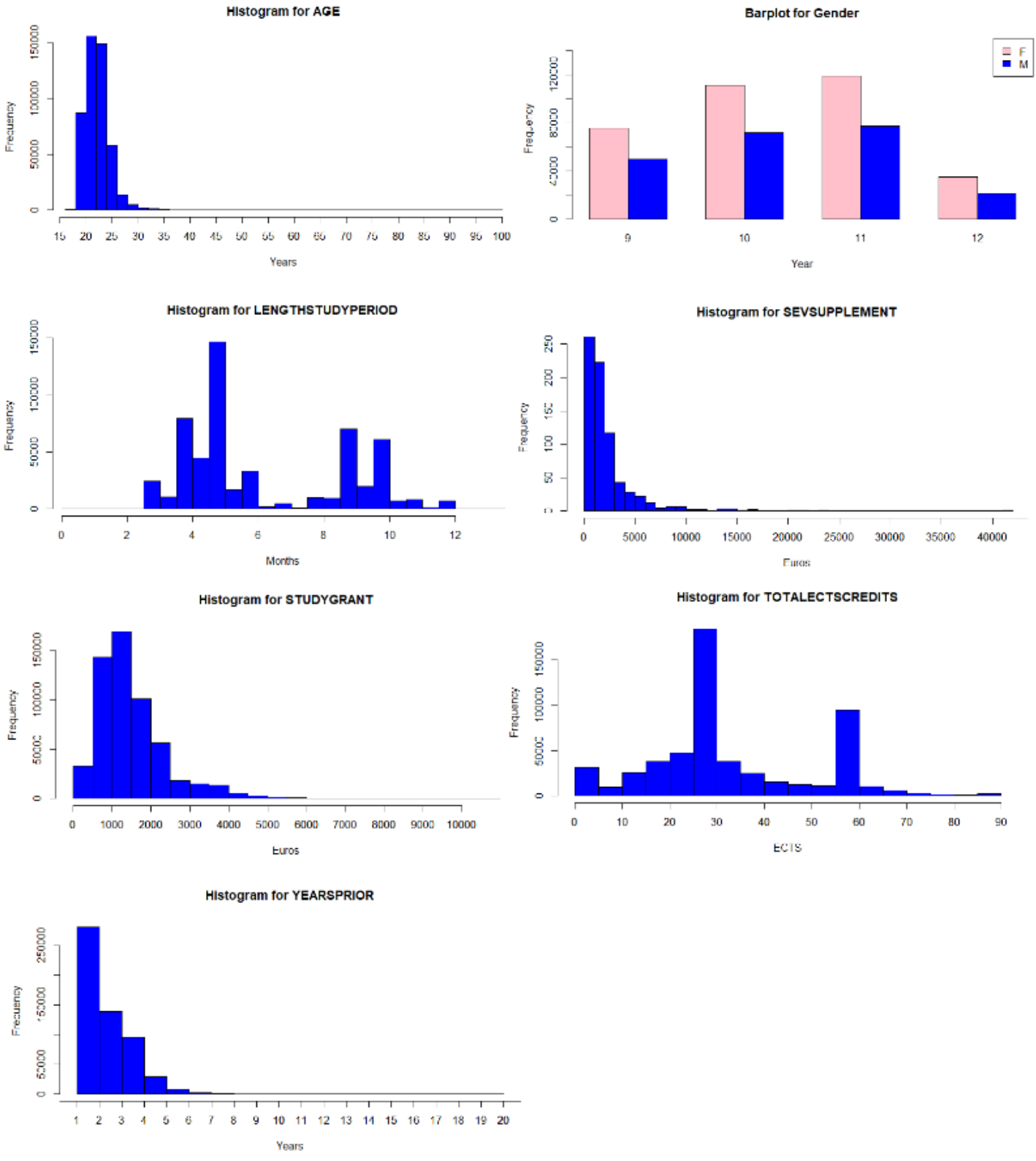


Figura 1. Histogramas e gráfico de barras para análise descritiva dos dados de Mobilidade de Estudantes. O valor máximo do eixo yy foi cortado em alguns histogramas de forma a ser possível visualizar-se uma gama de valores mais ampla no eixo dos xx.

4.2 Dados sobre o *ranking* das universidades

Através do *script* desenvolvido “rankingNotebook.Rmd” (ver ficheiro em anexo) foi efetuada uma análise descritiva dos dados relativos ao *ranking* das universidades em Rstudio. O resultado dessa análise para cada um dos anos considerados pode ser consultado nas figuras 24 E 25.

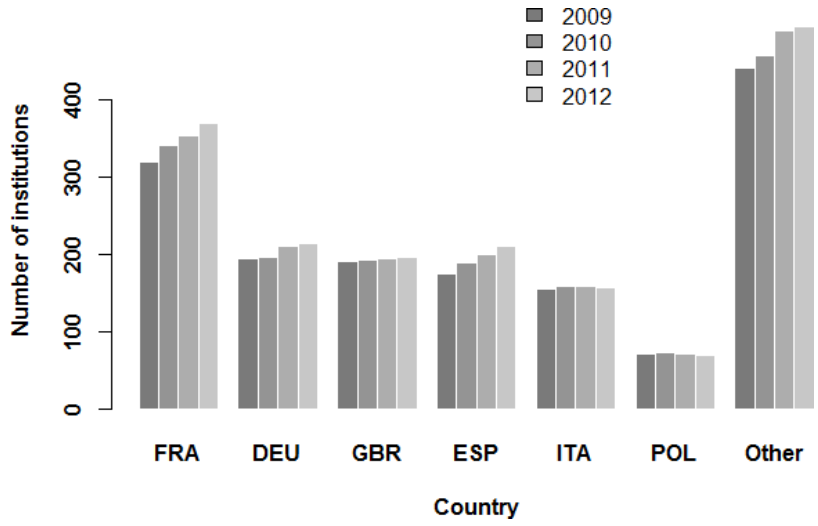


Figura 2. Países com maior número de instituições no *ranking*, para cada ano considerado.

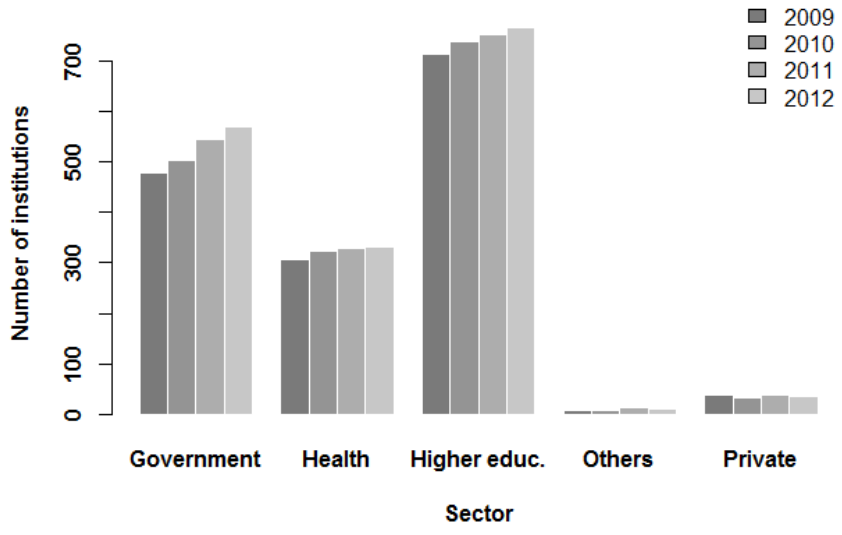


Figura 3: Número de instituições no *ranking* por sector, para cada ano considerado.

5. Diagrama

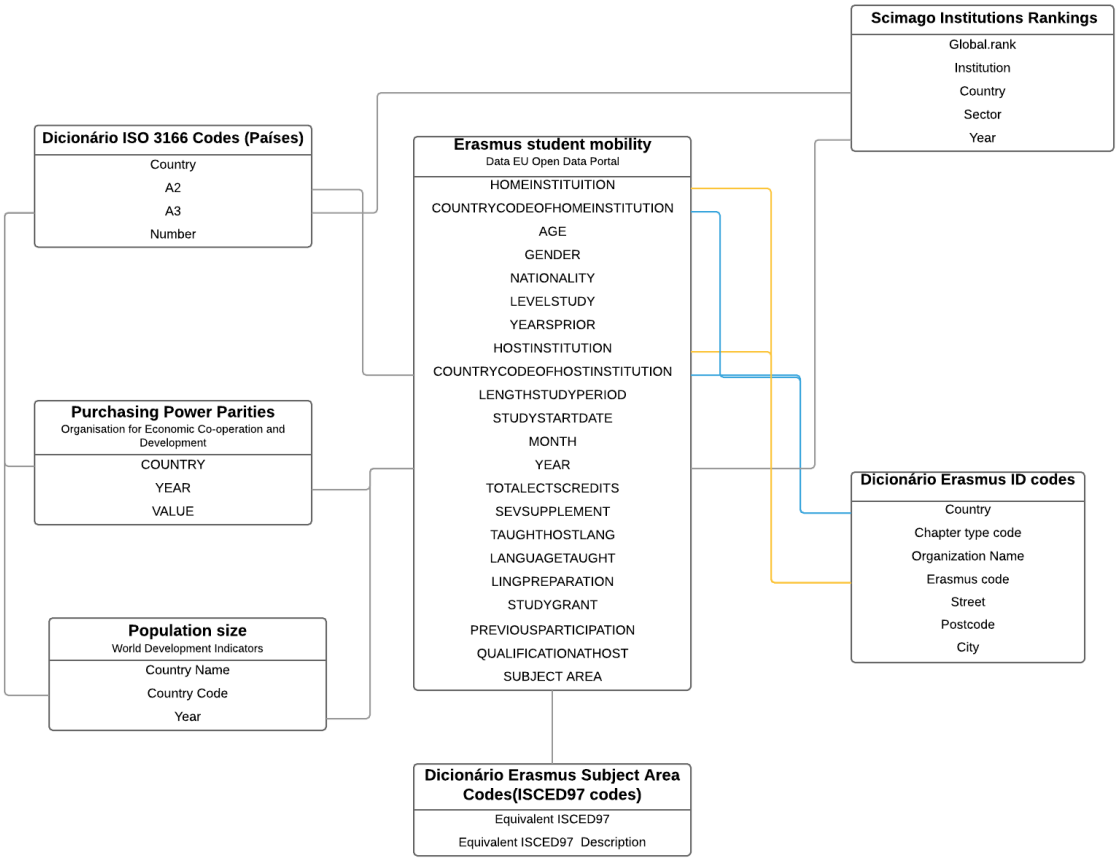


Figura 4: Diagrama com as ligações entre as fontes de dados.

6. Processo de negócio

O processo de negócio a modelar consiste na mobilidade de estudantes do Programa de Erasmus. Este programa é composto pelo estudante, a instituição e nsino superior que pertence o estudante e a instituição estrangeira que o acolhe. A duração da mobilidade é variável, com a duração mínima de 5 meses e máxima de 1 ano letivo. A cada estudante é concedida uma bolsa de mobilidade fixa para cobrir as despesas, nomeadamente as despesas de viagem e as decorrentes da diferença do custo de vida no país anfitrião. Os candidatos aprovados que não forem contemplados com bolsa poderão participar desde que haja vaga e suportam todas as despesas, mantendo-se com o estatuto de “Aluno Bolsa Zero”. Também existem bolsas especiais para estudantes portadores de deficiências. Neste contexto, irão ser identificados fatores impulsionadores da mobilidade, identificando parâmetros influenciadores, tais como a reputação da instituição e relevância das bolsas de mobilidade providenciadas, tendo em conta a duração da estadia e PPP do país anfitrião. Isto irá fornecer um suporte para futuras decisões em relação ao programa de mobilidade Erasmus.

7. Perguntas analíticas

- Como varia a soma total dos valores das bolsas atribuídas, tendo em conta o número de participantes, em cada ano académico?
- Quais são os países de origem com mais estudantes neste projecto?
- Quais as áreas de estudo em que foram efectuados mais ECTS?
- Como varia o número de participantes de acordo com o género?
- Qual o perfil dos estudantes portugueses?

8. Processo de negócio a modelar

O processo de negócio a modelar consiste na participação do estudante no programa Erasmus que envolve a atribuição de bolsas de mobilidade (em euros) pela Comissão Europeia, para a concretização de um certo número de ECTS (European credit transfer system).

Esta participação no programa tem um impacto ao nível dos desafios sociais, incluindo ações destinadas a promover a inclusão social e a garantir a aquisição pelos jovens de competências sociais, cívicas e interculturais, bem como de pensamento crítico. Este programa tem funcionado como um impulsionador do sentimento de união no contexto europeu, o que complementa e enriquece as entidades nacionais e regionais.

9. Grão e tipologia da tabela de factos

A modelação dimensional efectuada levou à criação da tabela de factos Student mobility, onde é possível consultar os dados necessários para analisar o processo de negócio indicado.

Grão: uma linha da tabela de factos corresponde à participação de um estudante (Student) no programa Erasmus, proveniente de uma instituição (Home institution), levada a cabo numa determinada instituição de acolhimento (Host institution), onde se encontra inscrito num determinado número de ECTS (Total ECTS credits), participação essa que começa numa determinada data (Date). A cada estudante é atribuído ou não uma bolsa por motivos de deficiência (Disability scholarship) e/ou uma bolsa de estudo (Study grant).

Tipologia: tabela de factos de transações que acumula participações dos estudantes no programa. Para a análise que pretendemos efectuar é suficiente saber apenas a data de início da participação, visto ser este o momento de atribuição das bolsas. Por outro lado, a actualização dos factos na tabela é inexistente, ou seja, uma vez ocorrido o evento numa determinada data, jamais haverá necessidade de modificar os seus dados, permitindo assim a consulta de dados históricos.

10. Dimensões

10.1. Dimensão Student (Quem?)

10.1.1. Descrição dos atributos

Tabela 9: descrição dos atributos da dimensão *Student.*

Atributo	Tipo de Dados	Descrição	Exemplo
Student key	Número inteiro	Chave substituta da dimensão Student. O seu domínio consiste num número inteiro que varia entre 1 e o número de linhas com dados.	1
Student natural key	Número inteiro	Chave supernatural da dimensão Student. Código do aluno. O seu domínio consiste num número inteiro que varia entre 1 e o número de linhas com dados.	1
Student age	Número inteiro	A idade do aluno definida como sendo a diferença entre o ano do início do Ano Erasmus, e o ano em que o aluno nasceu.	20
Student gender	Texto	Género do aluno. O seu domínio consiste em Male ou Female.	Male
Student nationality	Texto	Nacionalidade do aluno. O seu domínio consiste numa sequência de 2 caracteres no conjunto de "ISO Country Codes".	BE
Student years completed	Número inteiro	Número de anos completados de educação superior previamente ao período no estrangeiro.	1
Student previous participation	Texto	Reporta participação prévia no programa Erasmus. O seu domínio consiste numa das seguintes sequências de caracteres: No previous participation, Previous participation, Professional internship, Erasmus mundus	Erasmus mundus
Student row effective date	Data	Data de início da validade, no formato dd/mm/aaaa.	01/01/2009
Student row expiration date	Data	Data de fim da validade, no formato dd/mm/aaaa.	31/12/2009
Student current row indicator	Texto	Indicador que indica se a linha está em vigor. O seu domínio é uma sequência de caracteres, *Expired* ou *Current*.	Expired

10.1.2. Registo de mudanças lentas

A dimensão *Student* é uma dimensão de mudança lenta. Caso sejam feitas atualizações recorrer-se-á à técnica tipo 2. Esta técnica envolve uma chave supernatural, *Student natural key* e colunas extra referentes à validade (início e fim de validade) e a indicação se a linha está ou não em vigor.

10.2. Dimensão Institution (Onde?)

10.2.1. Descrição dos atributos

Tabela 10: descrição dos atributos da dimensão *Institution*.

Atributo	Tipo de Dados	Descrição	Exemplo
Institution key	Número inteiro	Chave substituta da dimensão Institution. O seu domínio consiste num número inteiro que varia entre 1 e o número de linhas com dados	100
Institution natural key	Número inteiro	Chave supernatural da dimensão Institution. O seu domínio consiste num número inteiro que varia entre 1 e o número de linhas com dados.	1
Institution Erasmus code	Texto	Chave supernatural da dimensão Institution. Código da instituição, segundo o conjunto de "Erasmus ID codes".	UK LONDON029
Institution name	Texto	Nome da instituição. O seu domínio é uma *string* de dimensão variável.	University College London
Institution country name	Texto	Nome do país da instituição (sequência de caracteres)	Portugal
Institution country A2 code	Texto	Código que identifica o país a que a instituição pertence. O código é constituído por uma string de duas letras correspondente ao respectivo código A2 do Dicionário ISO 3166 Codes para um dado país	AT
Institution country A3 code	Texto	Código que identifica o país a que a instituição pertence. O código é constituído por uma *string* de três letras correspondente ao respectivo código A3 do Dicionário ISO 3166 Codes para um dado país	GBR
Institution city	Texto	Nome da cidade da instituição (sequência de caracteres)	London
Institution street	Texto	Rua da instituição (sequência de caracteres).	Gower Street
Institution postcode	Texto	Código postal da instituição (sequência de caracteres).	WC1E 6BT
Institution country population	Número inteiro	Valor correspondente à população do país no ano de 2009. O seu domínio é um número inteiro entre 35766 e 81902307	81902307
Institution country PPP	Número decimal	Valor correspondente à paridade do poder de compra (PPP) do país no ano de 2009. O seu domínio é a sequência #N/A para países	1.5672

		sem informação ou um número com quatro casas decimais entre 0.469381 e 127.6870.	
Row effective date	Texto	Data de início da validade, no formato dd/mm/aaaa.	01/01/2009
Row expiration date	Texto	Data de fim da validade, no formato dd/mm/aaaa.	31/12/2009
Current row indicator	Texto	Indicador que indica se a linha está em vigor. O seu domínio é uma sequência de caracteres, *Expired* ou *Current*.	Expired

Embora inicialmente estivesse previsto a utilização de dados sobre Ranking das instituições, os mesmos não foram considerados nesta fase pois apenas um conjunto reduzido de instituições se encontrava incluído simultaneamente nos rankings e nos dados de mobilidade de estudantes.

10.2.2. Hierarquia

Na dimensão *Institution* existe a seguinte hierarquia de profundidade fixa: País > Cidade > Rua.

10.2.3. Registo de mudanças lentas

A dimensão *Institution* é uma dimensão de mudança lenta, uma vez que não é expectável que os atributos associados a uma determinada instituição mudem frequentemente. Para além disso, o surgimento de novas instituições também não é um evento comum. No entanto os atributos *Institution country PPP* e *Institution country population*, poderão variar anualmente e como tal recorrer-se-á à técnica tipo 2. Esta técnica envolve uma chave supernatural, *Institution natural key* e colunas extra referentes à validade (início e fim de validade) e a indicação se a linha está ou não em vigor.

10.3. Dimensão *Date* (Quando?)

10.3.1. Descrição dos atributos

Tabela 111: descrição dos atributos da dimensão *Date*.

Atributo	Tipo de Dados	Descrição	Exemplo
Date key	Número inteiro	Chave substituta da dimensão date. O seu domínio é um número inteiro entre 1 e o número de linhas presentes.	23
Date academic year	Texto	Ano académico a que se refere a mobilidade do estudante. O seu domínio é: 2009/2010, 2010/2011, 2011/2012.	2011/2012
Date year	Número inteiro	Ano do mês. O seu domínio é: 2009, 2010, 2011, 2012.	2011
Date semester	Número inteiro	Semestre do mês. O seu domínio é 1 ou 2.	1
Date month name	Texto	Nome do mês. O seu domínio é: Janeiro, Fevereiro, Março, Abril, Maio, Junho, Julho, Agosto, Setembro, Outubro, Novembro, Dezembro	Março
Date month number	Número inteiro	Número do mês da data. O seu domínio varia entre 1 e 12.	4

10.3.2. Hierarquia

Na dimensão *Date* existem duas hierarquias de profundidade fixa: Ano académico > Ano > Semestre > Mês (nome) e Ano académico > Ano > Semestre > Mês (número).

10.4. Dimensão Study Profile (Mini-dimensão)

10.4.1 Descrição dos atributos

Tabela 12: descrição dos atributos da dimensão *Study Profile*.

Atributo	Tipo de Dados	Descrição	Exemplo
Study profile key	Número	Chave substituta da mini-dimensão Study Profile. O seu domínio é um número inteiro entre 1 e o número de linhas presentes.	1
Subject area	Número inteiro	A área de estudo do estudante na sua instituição. É composto por um código designado ISCED97 codes ou Erasmus subject area codes.	225
Learning in host language	Texto	Reporta se a aprendizagem foi feita na língua do país anfitrião. O domínio consiste numa das seguintes sequências de caracteres: Taught in host language e Not taught in host language	Taught in host language
Language taught	Texto	A linguagem na qual se efetuou a aprendizagem. O seu domínio consiste numa sequência de dois	EN

		caracteres, de acordo com o conjunto “ISO Language Codes”.	
Language preparation	Texto	Reporta se o estudante frequentou um curso de línguas no país anfitrião ou de origem, ou um outro tipo de curso relacionado com Erasmus. O seu domínio consiste numa das seguintes sequências de caracteres: EILC preparation, Host language preparation, Home language preparation ou No preparation	Home language preparation
Qualification at host	Texto	Reporta a qualificação que o estudante receberá na instituição anfitriã. O seu domínio consiste numa das seguintes sequências de caracteres: double diploma, joint diploma, other qualification, no qualification	No qualification
Student level of study	Texto	Grau dos estudos na instituição de origem (Ciclo). O seu domínio consiste numa das seguintes sequências de caracteres: First cycle, Second cycle, Third cycle, Short cycle	First cycle

Esta mini-dimensão referente ao perfil de estudo acolheu atributos da dimensão *Student* para evitar que se tornasse numa dimensão “monstra”.

11. Estrutura da tabela de factos

11.1. Atributos da tabela de factos

Tabela 13: descrição dos atributos da tabela de factos.

Atributo	Tipo de Dados	Descrição	Exemplo
Student key	Número inteiro	Chave estrangeira para a dimensão Student. Corresponde ao atributo Key da dimensão Student (chave substituta). O seu domínio consiste num número inteiro que varia entre 1 e o número de linhas com dados.	2
Student profile key	Número inteiro	Chave estrangeira para a dimensão Student Profile Key. Corresponde ao atributo Key da dimensão Study Profile (chave substituta). O seu domínio consiste num número inteiro que varia entre 1 e o número de linhas com dados.	1
Date key	Número inteiro	Referencia a chave estrangeira para a dimensão Date. O seu domínio é um número inteiro entre 1 e o número de linhas presentes.	25
Home institution	Número inteiro	Chave estrangeira para a dimensão Institution, relativa à instituição de origem. Corresponde à chave substituta dessa dimensão, o atributo Institution Erasmus code..	37
Host institution	Número inteiro	Chave estrangeira para a dimensão Institution, relativa à instituição de acolhimento. Corresponde à chave substituta dessa dimensão, o atributo Institution Erasmus code.	434

A chave primária é composta pelas quatro chaves estrangeiras para as dimensões referidas na tabela 11. Isto significa que cada facto presente na tabela é identificado univocamente pelos valores das chaves estrangeiras referidas.
O valor de *Home institution* nunca é igual ao valor de *Host institution*, uma vez que, quando um estudante participa no programa, a participação ocorre sempre numa instituição diferente da instituição de origem.

11.2. Medidas numéricas da tabela de factos

- **Length study period:** é uma medida semi-aditiva porque não faz sentido somar o número de meses da duração da mobilidade, contudo é útil saber a duração média da mobilidade erasmus.

- **Disability scholarship** e **Study grant**: são medidas aditivas porque permitem a

soma de valores ao longo de todas as dimensões.

- **Total ECTS credits**: é uma medida semi-aditiva, pois não se obtém informação relevante ao somar créditos de acordo com todas as dimensões da tabela de factos. Contudo, a sua soma total é útil para responder a questões como qual as Áreas de Estudo em que os alunos completaram mais créditos.

Tabela 14: descrição das medidas da tabela de factos

Medida	Descrição	Tipo de Dados	Exemplo
Length of study period	Número decimal	Duração do período Erasmus em meses. O seu domínio consiste num valor numérico, com duas casas decimais, entre 0.00 e 12.00.	3.5
Total ECTS credits	Número inteiro	Número antecipado de ECTS feitos pelo estudante durante o período Erasmus. O seu domínio consiste num inteiro entre 0 e 90.	30
Disability scholarship	Número decimal	Bolsa concedida por motivos de deficiência. O seu domínio consiste num valor positivo da moeda (euro), com até duas casas decimais.	0.00
Study grant	Número decimal	Bolsa recebida pelo estudante, excluindo a bolsa de deficiência. O seu domínio consiste num valor positivo da moeda (euro), com até duas casas decimais.	1780.03

12. Diagrama em estrela do *data warehouse*:

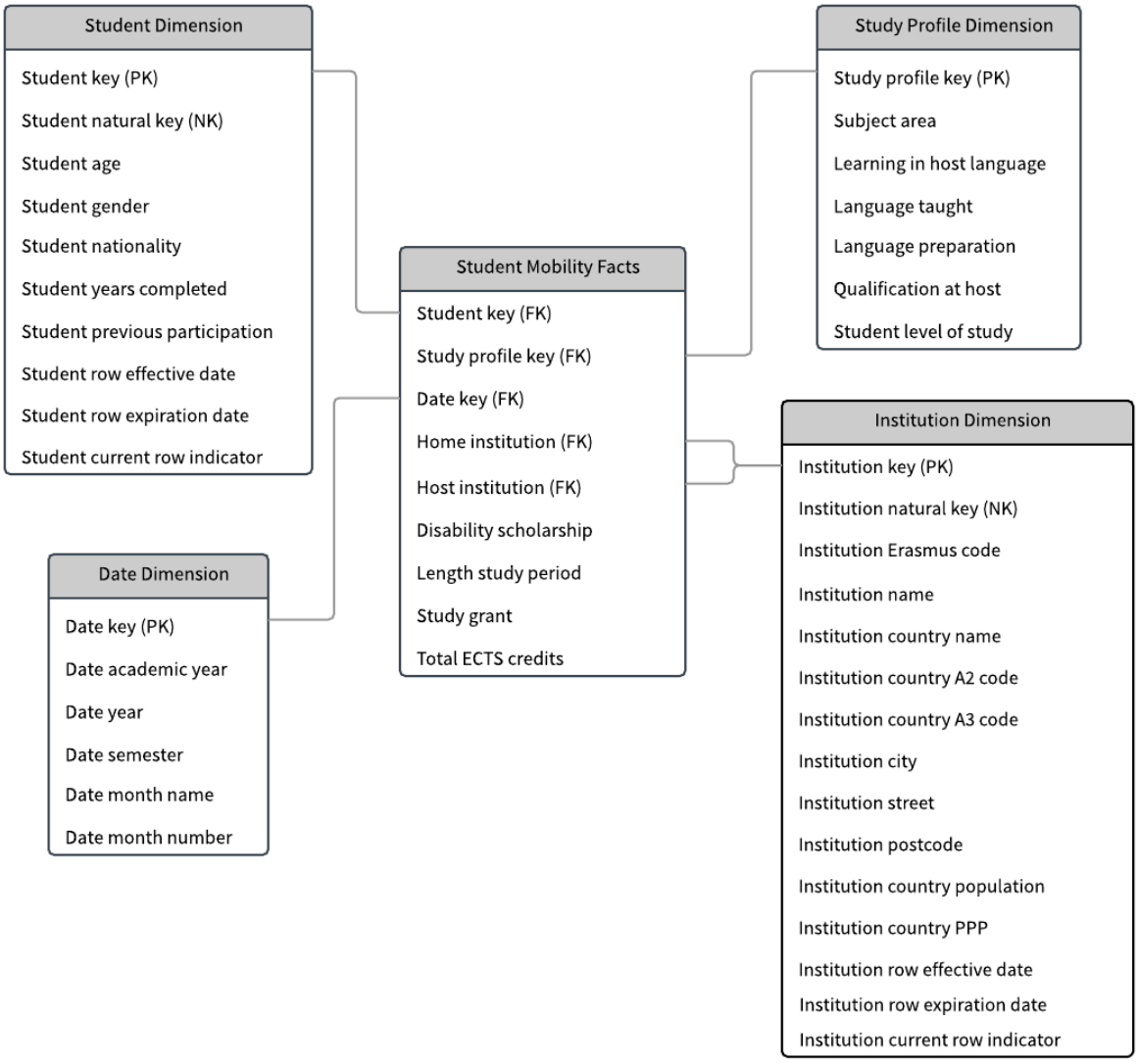


Figura 7: Diagrama em estrela do *data warehouse*.

13. Sistema ETL

13.1 Responsabilidades do sistema ETL

- **Extração (extraction):** Obtenção dos dados a partir dos repositórios de dados abertos previamente mencionados. Leitura, compreensão e selecção dos dados necessários para o *data warehouse*, extraídos para o *data staging area* de forma a permitir a sua manipulação. Esta fase foi feita manualmente. Os ficheiros obtidos foram os seguintes: '2009-2010.csv', '2010-2011.csv', '2011-2012.csv', 'Population_size.xls', 'PPP.csv', 'ISO 3166 Country Codes.tsv', 'Erasmus_codes.csv'

- **Transformação (transformation):** A etapa Transformation do sistema ETL é levada a cabo na *data staging area*. Consiste na limpeza dos dados, combinação de múltiplas fontes e atribuição de chaves primárias e chaves supernaturais. Preparação da estrutura que possibilita a aplicação de *queries* e análise (Dimensões e Tabela de Factos) – modelação dimensional.
  Para efectuar esta fase, foram utilizados múltiplos scripts em Python.

  Scripts Python de Limpeza

  read_write.py, recebe um ficheiro csv e retorna um leitor, que consiste numa lista de dicionários, em que cada dicionário é uma linha do ficheiro csv com as respectivas colunas na chave de cada item.

  Cleaning.py, recebe o ficheiro csv dos dados de Mobilidade de Estudantes ('Student_mobility_raw.csv') e efectua as tarefas de limpeza previamente mencionadas no ponto 3.1, originando um novo ficheiro csv ('Student_mobility_clean.csv')

  Cleaning_Erasmus_codes.py, recebe o ficheiro csv dos códigos Erasmus das instituições participantes ('Erasmus_codes.csv') e efectua a limpeza conforme mencionado no ponto 3.1, originando um novo ficheiro csv

('Erasmus_code_clean.csv').

Scripts Python de Modelação dimensional

modelling.py, chama os *scripts* responsáveis por modelar as dimensões a partir do ficheiro de entrada 'Student_mobility_clean.csv', nomeadamente, dimStudent.py, dimStudyProfile.py, dimInstitution.py, dimDate.py. O *script* dimInstitution.py recebe ainda como parâmetro os dicionários 'subject_areas.csv', ''ISO 3166 Country Codes.tsv', 'Population_size.xls' e 'PPP_clean_data.csv'.
Os ficheiros de saída são os seguintes: 'dimStudent.csv', 'dimStudyProfile.csv', 'dimInstitution.csv', 'dimDate.csv'

Criação da tabela de factos

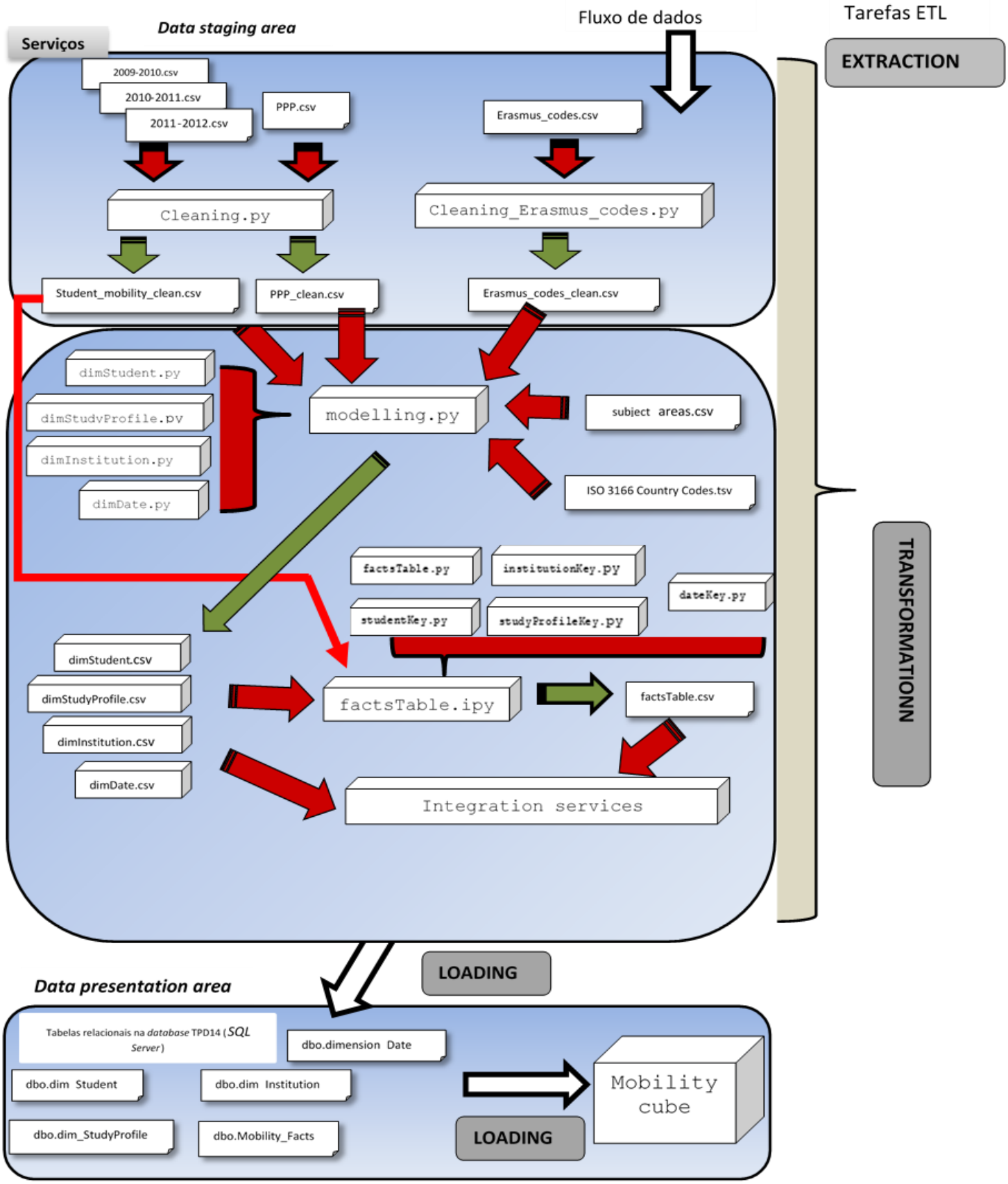
IPython notebook **factsTable.ipynb**:

- **Entradas**: necessita dos seguintes ficheiros na mesma directoria: studentKey.py, studyProfileKey.py, institutionKey.py, dateKey.py, measures.py, factsTable.py, 'Student_mobility_clean.csv', 'dimStudent.csv', 'dimStudyProfile.csv', 'dimInstitution.csv', 'dimDate.csv'
- **Tarefas**: importa e executa o Python script factsTable.py, que por sua vez importa e executa os scripts studentKey.py, studyProfileKey.py, institutionKey.py, dateKey.py, measures.py que extraem, respectivamente, as chaves substitutas de cada dimensão de acordo com a ordem dos dados presentes em 'Student_mobility_clean.csv' e as medidas 'Lenght of study period', 'Total ECTS', 'Disability scholarship' e 'Study scholarship'. É também construída a tabela de factos contendo colunas com as chaves estrangeiras de cada uma das dimensões, bem como as medidas referidas.
- **Saída**: 'facts_table.csv', contendo as colunas 'Student Key', 'Study profile key', 'Home institution key', 'Host institution key', 'Date key', 'Lenght of study period', 'Total ECTS', 'Disability scholarship' e 'Study scholarship'.

- **Carregamento (loading):** Carregamento dos dados a partir de ficheiros CSV para o *SQL Server*, com um projeto de *Integration Services* do *SQL Server Data Tools*. Obtenção das tabelas relacionais para as dimensões 'Student', 'Study profile', 'Institution' e 'Date' na base de dados TPD14 (presente no servidor CACILHEIRO) e importação dos respectivos dados a partir dos ficheiros 'dimStudent.csv', 'dimStudyProfile.csv', 'dimInstitution.csv' e 'dimDate.csv'. Para além disso, cria a tabela relacional para a tabela de factos 'Student mobility facts' e importa os respectivos dados a partir do ficheiro 'facts_table.csv'.

  Implementação do cubo de dados através de um projecto de Analysis Services **'Mobility cube.sln'**, cuja entrada consiste nas tabelas relacionais presentas na base de dados 'TPD14' no servidor CACILHEIRO:
  dimension_InstitutionCorrected, dimension_Date, dimension_Student, dimension_StudyProfile, Mobility_factsCorrected.
  Criação do ficheiro '**TPD source.ds'**, criação de um ficheiro com a vista '**TPD view.dsv'**, criação dos ficheiros das dimensões 'Student.dim', 'Study Profile.dim', 'Institution.dim', 'Date.dim', definição das hierarquias presentes nas dimensões 'Institution' e 'Date', alteração dos nomes dos atributos para formas mais explícitas, criação do cubo de dados 'Mobility Cube'.
  Saída: cubo de dados '**Mobility Cube**' preparado para o 'Deployment'. A realização de relatórios dinâmicos através de ferramentas como Power BI Desktop ou Excel passa a ser possível após a implementação do cubo de dados.

13.2 Diagrama com fluxos de dados e programas do sistema ETL



14. Cubo de dados implementado

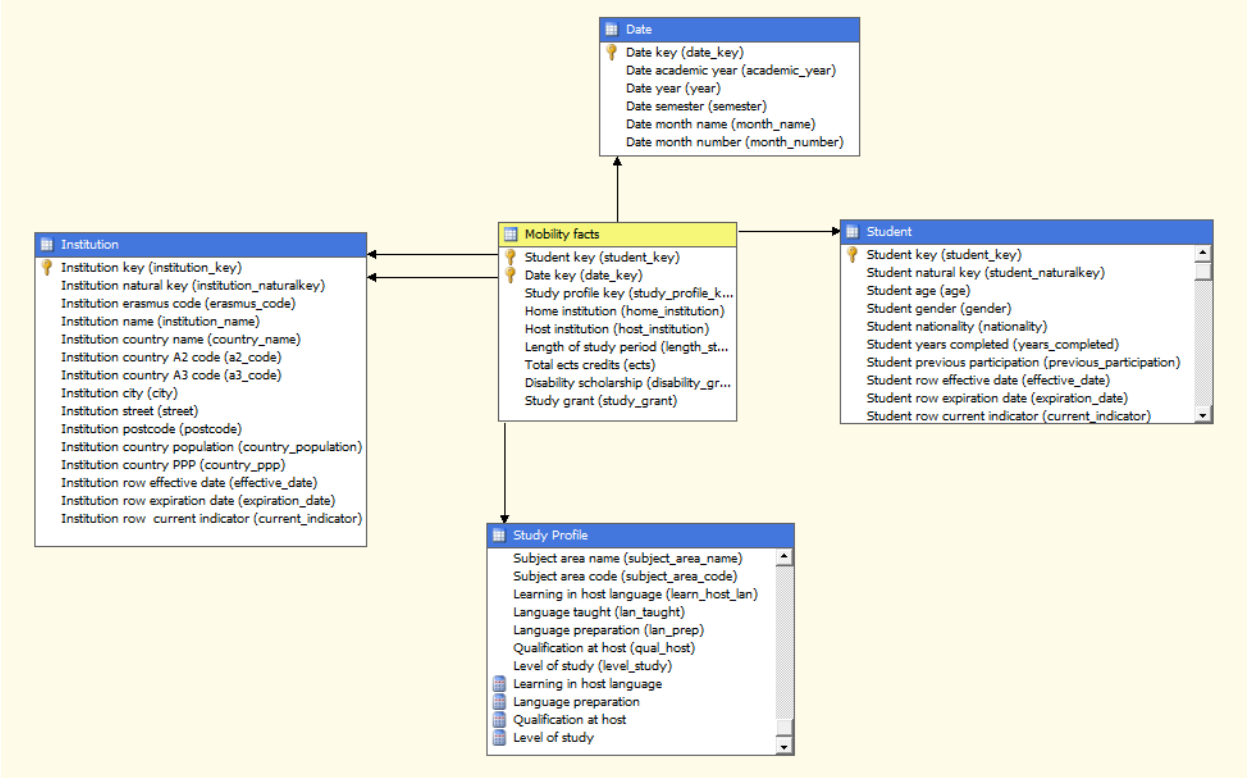


Figura 8: Dimensões e tabela de factos do cubo de dados implementado, nomeadamente, as dimensões Instituição, Data, Estudante, Perfil de Estudo, e, a tabela de factos Mobilidade.

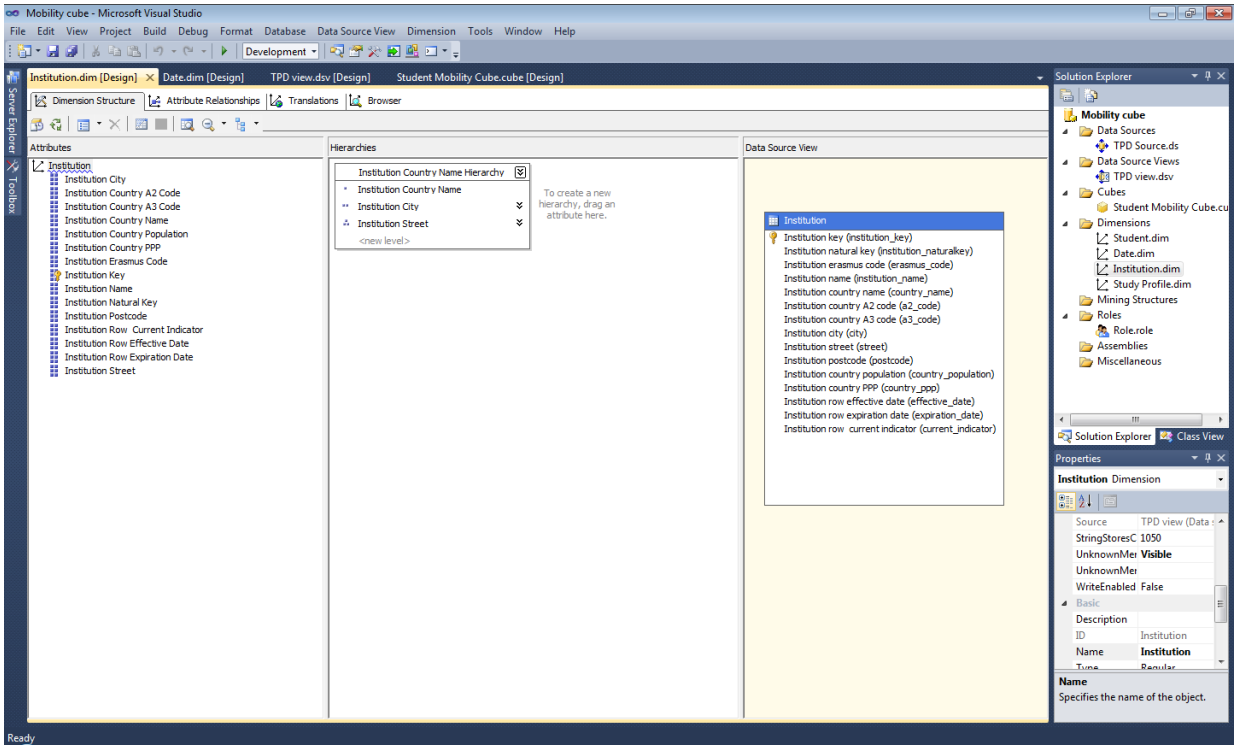


Figura 9: Hierarquia implementada na dimensão *Institution*.

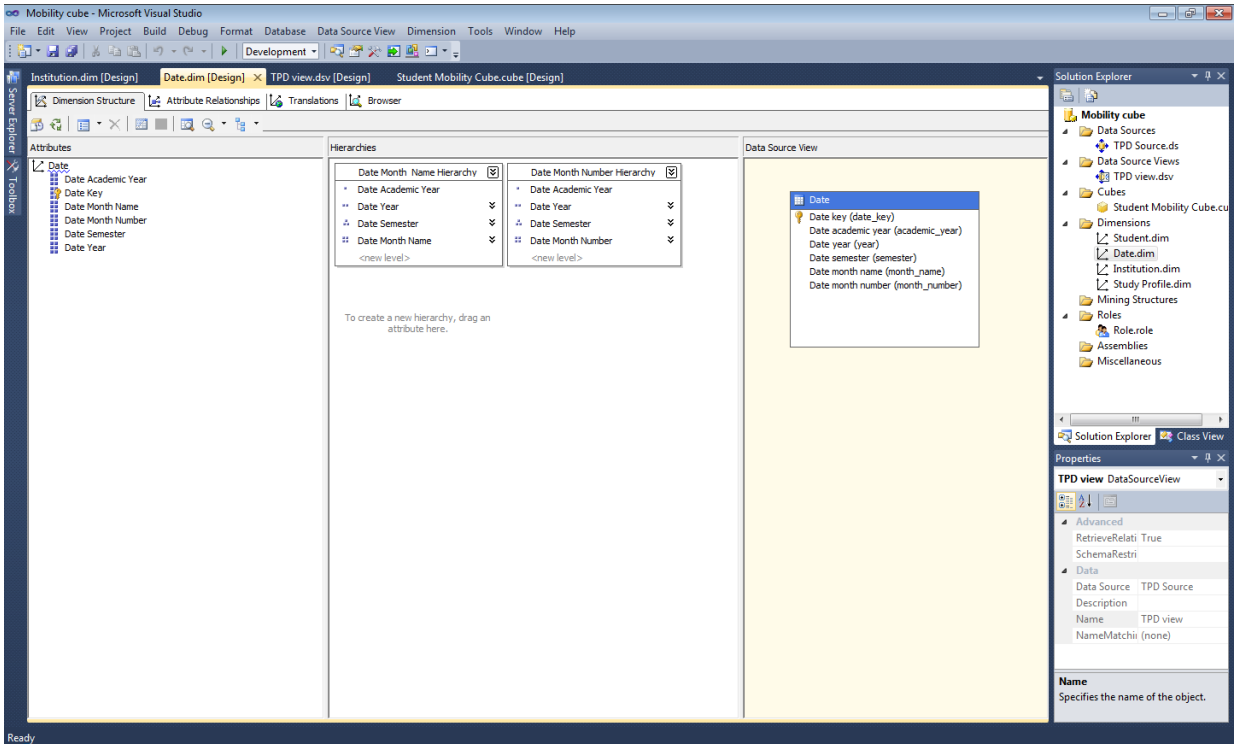


Figura 10: Hierarquia implementada na dimensão *Date.*

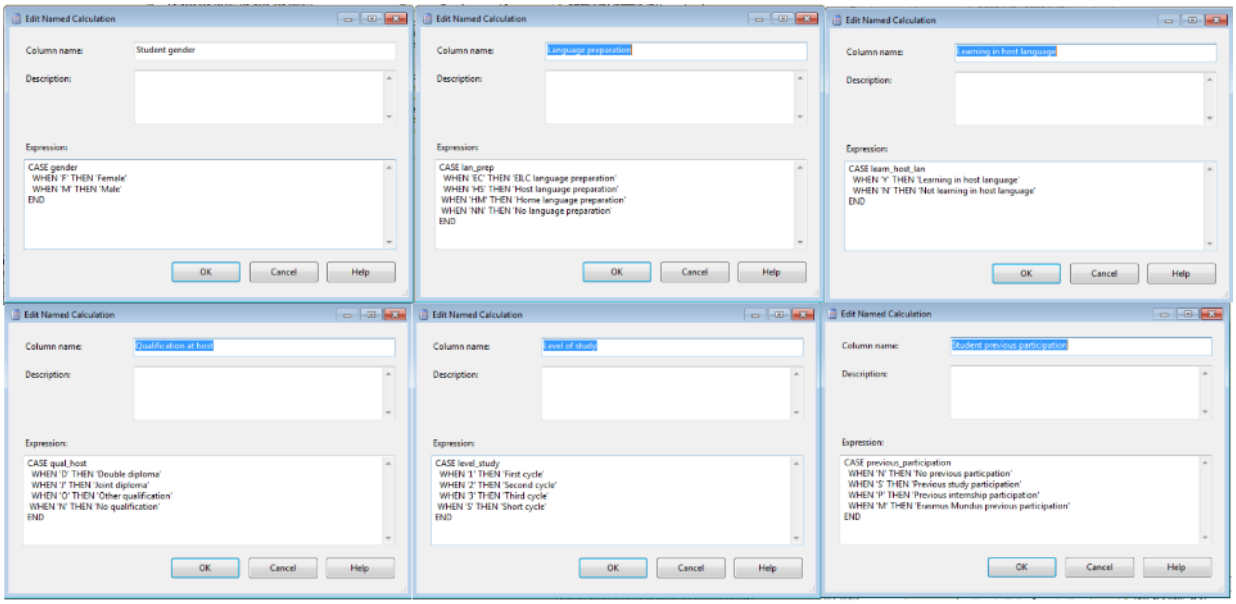


Figura 11: *Named Calculations* implementadas na dimensão *Student* (primeira) e *Study Profile* (restantes) de forma a aumentar a legibilidade da informação .

15. Relatórios analíticos

15.1. Como variam os valores das bolsas atribuídas, tendo em conta o número de participantes, em cada ano académico?

Os relatórios analíticos desta secção foram gerados através da ferramenta Excel. Na Figura 12 é possível observar a evolução do número de participantes ao longos dos anos académicos 2008/2009, 2009/2010, 2010/2011 e 2011/2012.

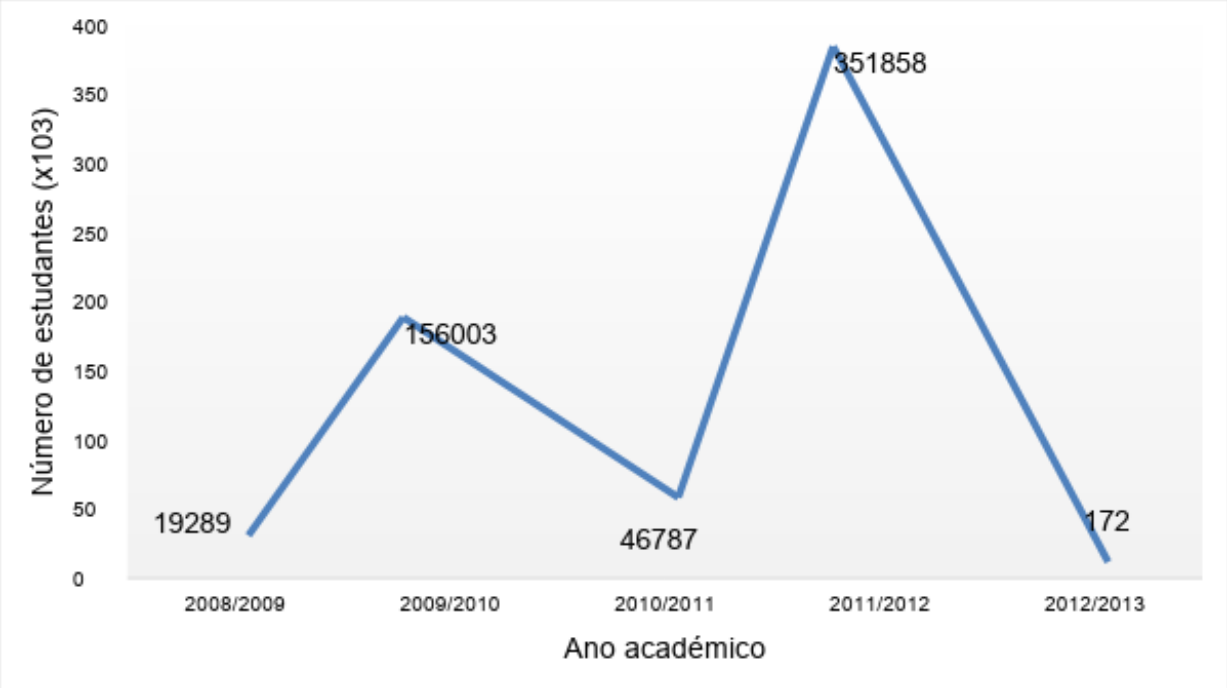


Figura 12: 'Line chart' representando a evolução do número de participantes (em milhares) por ano académico.

O número de participantes atinge o seu pico mínimo (19 289) no primeiro ano considerado (2008/2009) e o seu pico máximo (351 858) no último ano (2011/2012). No entanto, a evolução ao longo dos anos não é sempre crescente, pois o número de participantes em 2010/2011 é menor do que em 2009/2010. Existem apenas registos residuais para o ano de 2012/2013, tendo em conta que não foram obtidos dados relativos a este ano lectivo, sendo que os registos constavam no ficheiro relativo ao ano lectivo anterior.

A informação retirada do presente relatório parece sugerir uma tendência que à primeira vista se afigura contra-inituitiva, visto que a evolução do número de participantes é demasiado discrepante de um ano para o outro, em particular do ano 2009/2010 para o ano 2010/2011. Como tal, procedeu-se a uma análise complementar

de modo a descobrir as possíveis causas por detrás desta disparidade.

Para começar, foram elencadas várias hipóteses que possivelmente poderiam explicar esta situação:

1. Os ficheiros iniciais disponibilizados na fonte original apresentam um número de registos muito diferente para os vários anos académicos.
2. O processamento dos dados poderá ter eliminado desproporcionalmente mais registos respectivos ao ano lectivo 2010/2011 do que aos restantes anos.
3. Os registos constantes nos ficheiros não correspondem, necessariamente, ao ano lectivo do ficheiro em que aparecem ou existem dados omissos, logo, os ficheiros contêm dados irregulares ou incompletos.
4. Poderá ter havido uma diminuição de intercâmbios no ano lectivo em questão.

A comparação dos três ficheiros iniciais (relativos a 2009, 2010, e, 2011 e 2012) indica que existe um ligeiro aumento de participantes ao longo dos ficheiros (Anexo 1). A limpeza dos dados elimina proporcionalmente as linhas – quanto maior o número de participantes, maior o número de linhas eliminadas. Não se encontram discrepâncias suficientes entre os ficheiros originais para explicar o valor reduzido do ano lectivo 2010/2011. Os valores dos ficheiros pré-limpeza e pós-limpeza são, em geral, semelhantes, ocorrendo uma ligeira diminuição do número de linhas no ficheiro pós-limpeza. Conclui-se que a discrepância de participantes no ano lectivo em causa não tem como factor a limpeza dos dados. Contudo, tendo os 3 ficheiros iniciais em conta, verifica-se que, independentemente da limpeza, o número de participantes nos ficheiros originais no ano 2010 foi muito inferior ao ano anterior.

Além disso, verifica-se que existe uma variedade muito pequena de meses que podem ser enquadrados no ano lectivo 2010/2011 (assinalados a vermelho no anexo 2), comparativamente aos restantes anos lectivos. Verifica-se um padrão: os meses janeiro e fevereiro, setembro e agosto geralmente têm maior número de participantes. Contudo, os meses janeiro e fevereiro de 2011 não constam nos dados obtidos, o que poderá explicar a observação do número reduzido de participantes neste ano lectivo.

Assim, conclui-se que o número reduzido de participantes no ano lectivo 2010/2011 não está relacionado com o tratamento de dados, excluindo-se, por conseguinte, a hipótese 2. Logo, o motivo poderá ser, efectivamente, uma ausência de intercâmbios nesse período específico (hipótese 4) ou ainda a falta de registo e/ou

disponibilidade dos registos nos dados providenciados no portal (hipótese 3).

De seguida, efectuou-se uma análise do valor da bolsa de estudo atribuída por participante em cada ano académico, cujo resultado é possível observar na figura 13. O ano académico 2012/2013, dado o reduzido número de registos existentes, foi excluído da análise de forma a evitar o enviesamento da média calculada para todos os anos em conjunto.

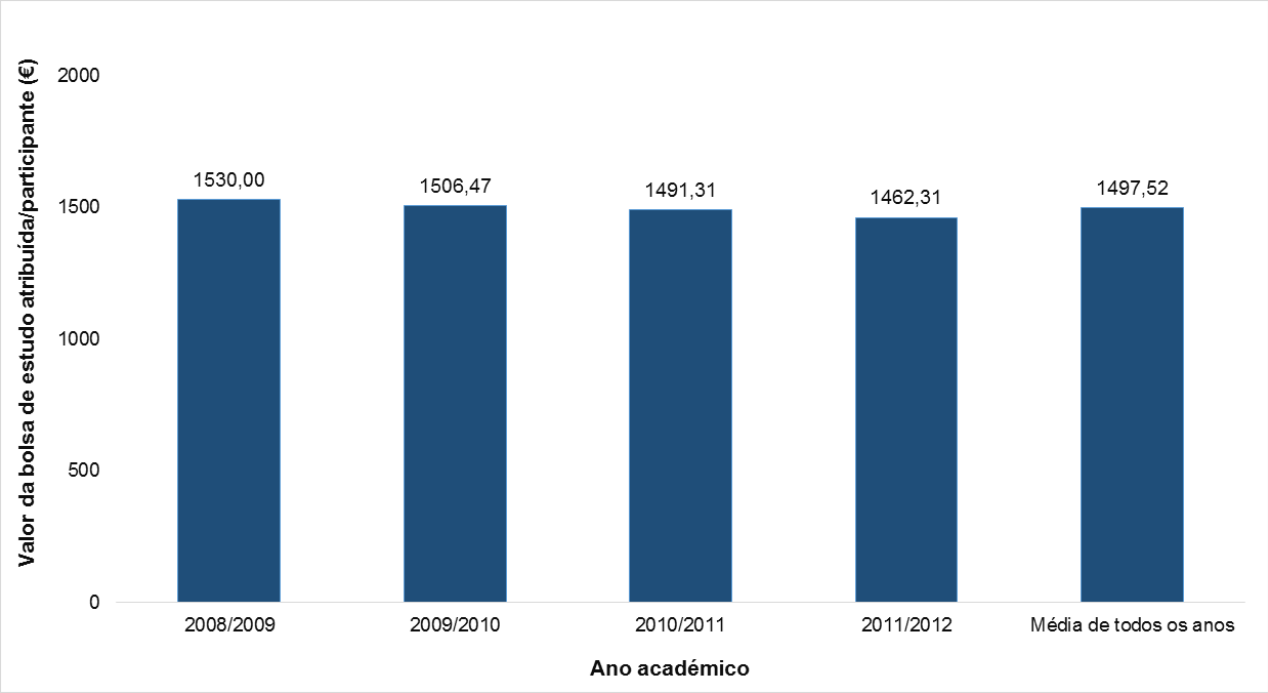


Figura 13: 'Bar chart' com o valor da bolsa atríbuida por participante com bolsa de estudo para os anos académicos 2008/2009, 2009/2010, 2010/2011 e 2011/2012 e o valor médio de todos os anos considerados em conjunto.

O valor da bolsa atríbuida por participante manteve-se estável ao longo dos anos em estudo, variando entre um mínimo de 1462,31€ no ano 2011/2012 e um máximo de 1530,00€ no ano 2008/2009. Ambos os valores não distam de forma acentuada da média de todos os anos, cujo valor é 1479,13€.

A análise foi refinada através da obtenção do valor da bolsa atribuída/participante bolseiro para cada país que recebeu alunos em cada um dos anos referidos.

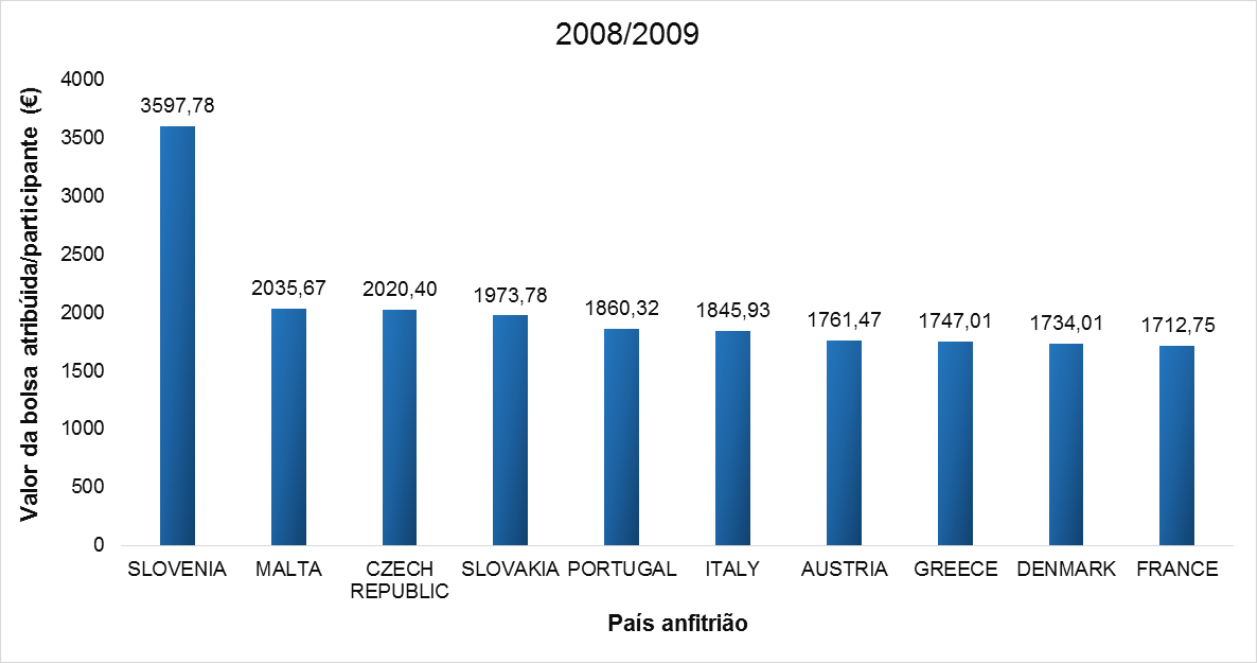


Figura 14: 'Bar chart' representando os dez países anfitriões com o maior valor da bolsa atríbuida/participante com bolsa de estudo para o ano académico 2008/2009.

No ano 2008/2009 a lista de países é liderada pela Eslovénia (Figura 14), em que o valor da respectiva bolsa por participante é 3597,78 €. Este valor encontra-se bastante acima do valor médio do ano académico referido, que é 1530,0 €. O segundo país da lista, Malta, apresenta um valor bastante mais reduzido (2035,7 €). A lista é liderada por dois dos países com menor população, o que sugere um possível enviesamento nos valores obtidos em detrimento dos países mais populosos.

Destaca-se também o facto de os restantes países que compõe a lista apresentarem, sem excepção, valores acima do valor médio do ano académico. A explicação poderá advir do reduzido número de registos disponíveis neste ano que enviesam os valores obtidos.

Na figura 15, observa-se a mesma análise efectuada para o ano académico 2009/2010

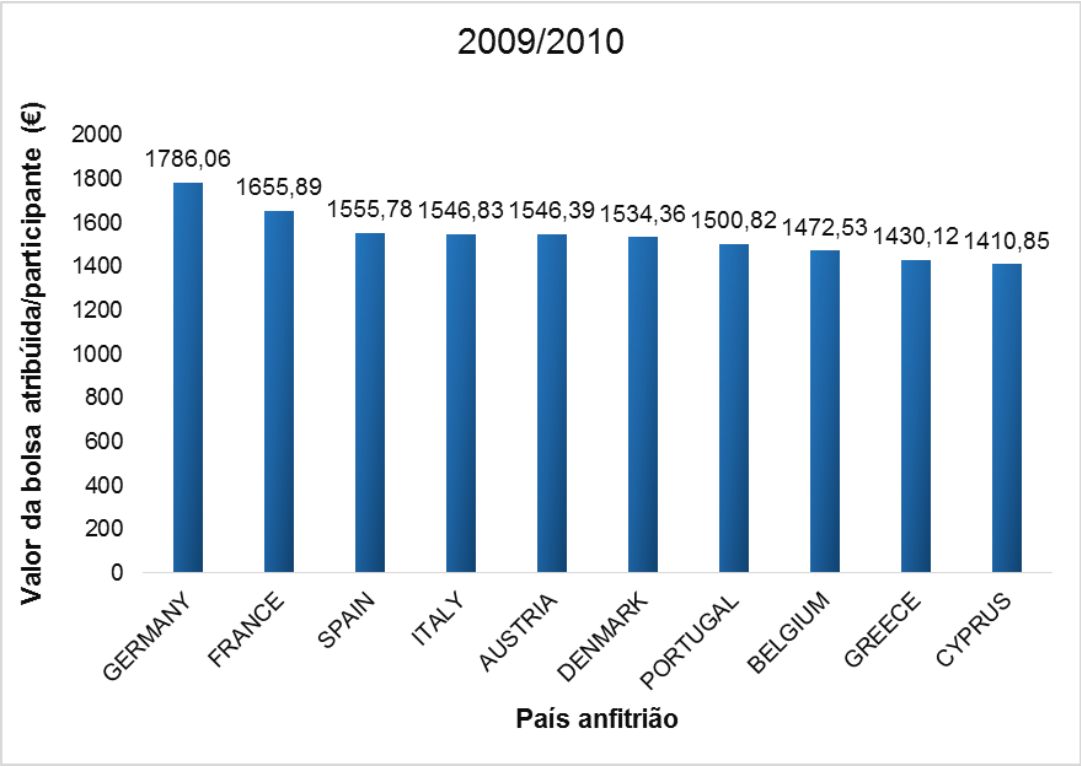


Figura 15. 'Bar chart' representando os dez países anfitriões com o maior valor da bolsa atribuída/participante com bolsa de estudo para o ano académico 2009/2010.

No ano 2009/2010, a lista dos países anfitriões é liderada pelos países mais populosos: Alemanha, França e Espannha, com os valores 1786,06, 1655,89 e 1555,78 euros.

Na figura 16, a análise efectuada para o ano 2010/2011 pode se consultada. A República Checa apresenta o valor mais elevado atribuido por participante 1944,05 euros.

Por último, na figura 17 encontram-se os dados relativos ao ano 2011/2012, cuja lista de países é novamente liderada por um país pouco populoso, Liechtenstein. De destacar que este país é seguido pela Alemanha, França e Espanha que ocupam a segunda, a terceira e a quarta posição. Estes três países compunham o pódio para o ano 2009/2010

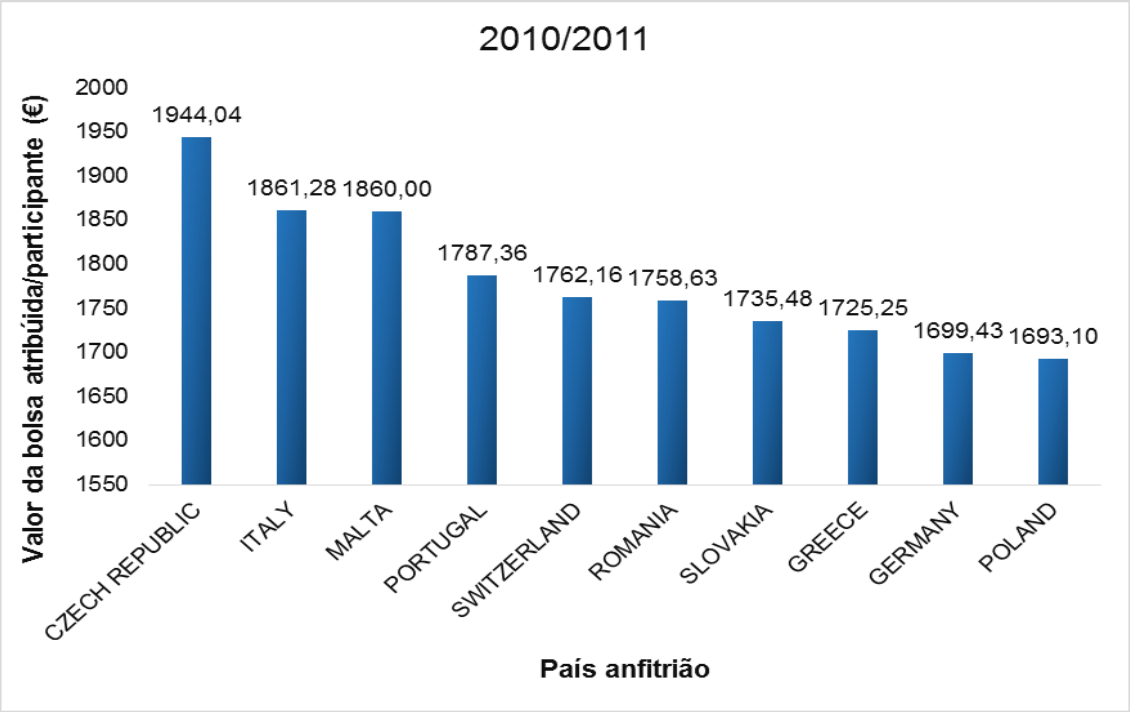


Figura 16: 'Bar chart' representando os dez países anfitriões com o maior valor da bolsa atríbuida/participante com bolsa de estudo para o ano académico 2010/2011.

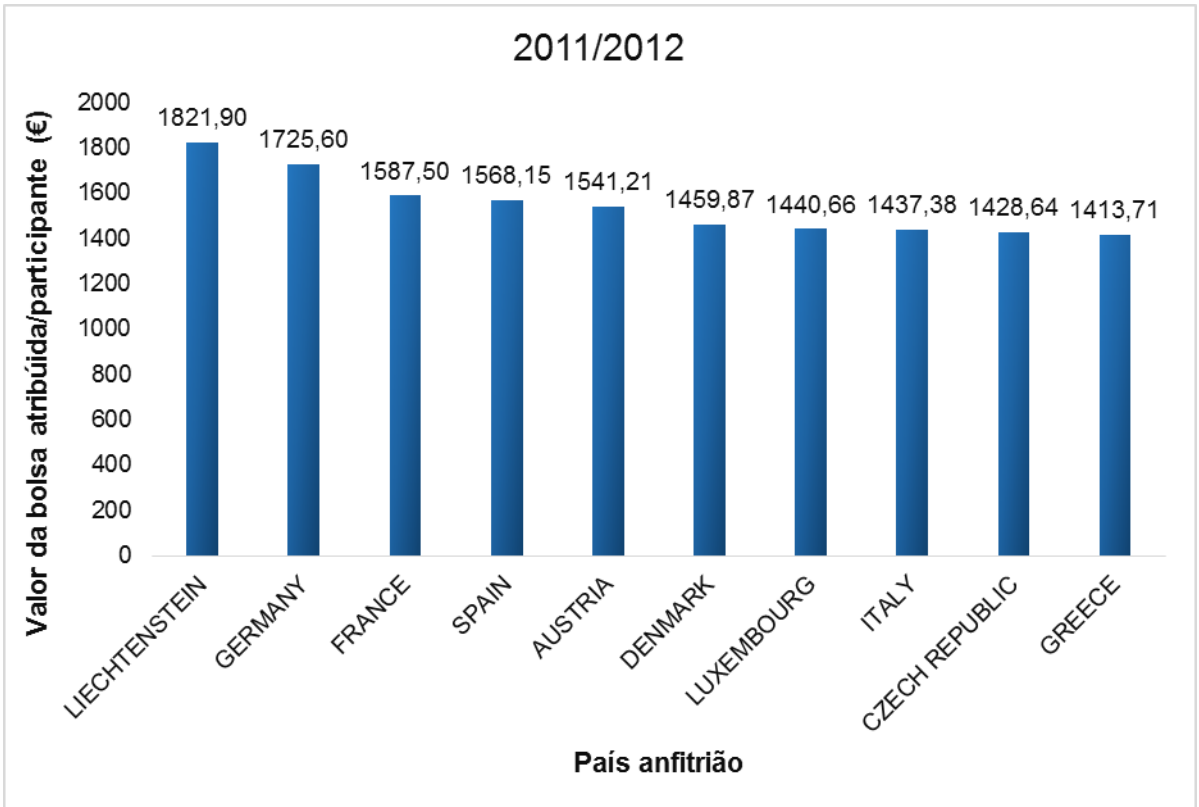


Figura 17: 'Bar chart' representando os dez países anfitriões com o maior valor da bolsa

atríbuida/participante com bolsa de estudo para o ano académico 2011/2012.

15.2. Quais são os países de origem com mais estudantes neste projecto?

Na Figura 18 é possível observar todos os países de origem dos participantes do programa Erasmus. É de salientar o grande número de países Europeus em que os estudantes optam por este programa de educação.

Figura 18: 'Filled map' representando os países de origem dos estudantes do programa Erasmus

Na Figura 19 é possível observar os dez países com mais participantes no programa Erasmus em termos de números brutos. Destacam-se Espanha e Alemanha como os países com mais participantes, seguidos pela França. As últimas posições do *top* são ocupadas por Portugal, República Checa, Países Baixos e Bélgica com número semelhante de participantes.

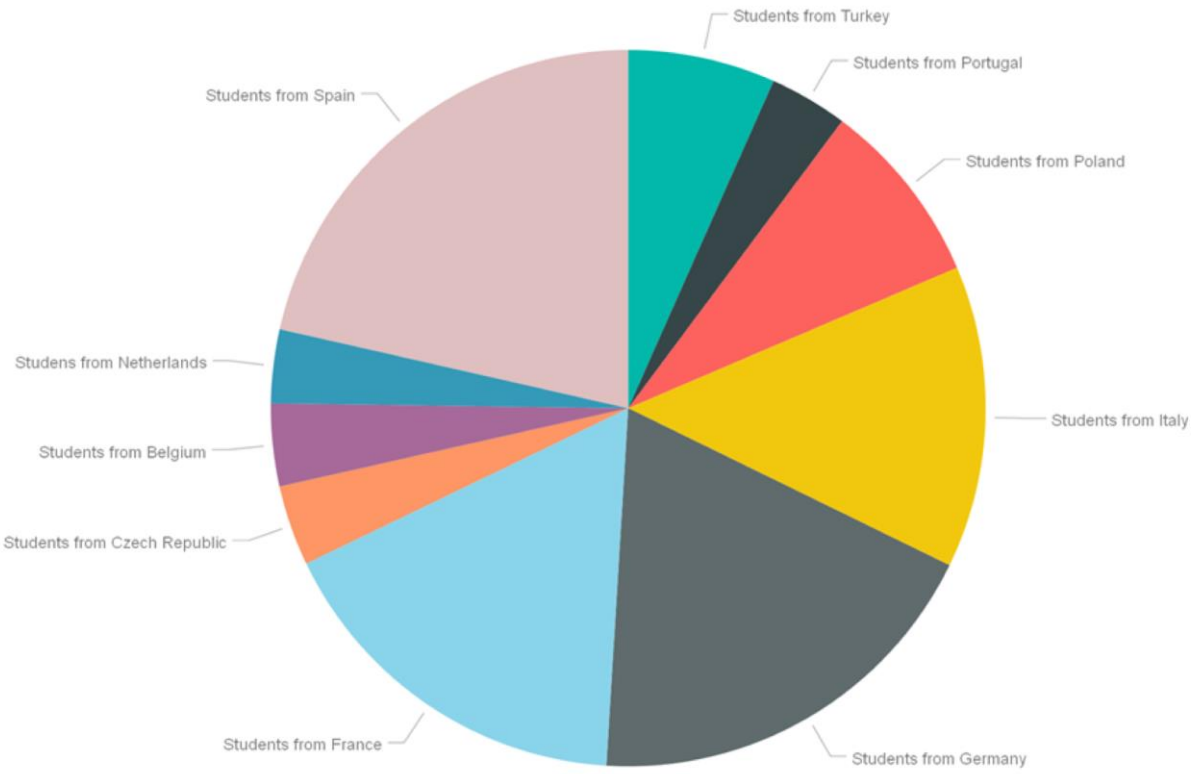


Figura 19: 'Piechart' com o *top* 10 dos países de origem dos estudantes do programa Erasmus.

Na Figura 20 é fácil assumir de que há um maior número de estudantes em Erasmus quanto maior a população do país em questão, sendo a Turquia o caso de exceção.

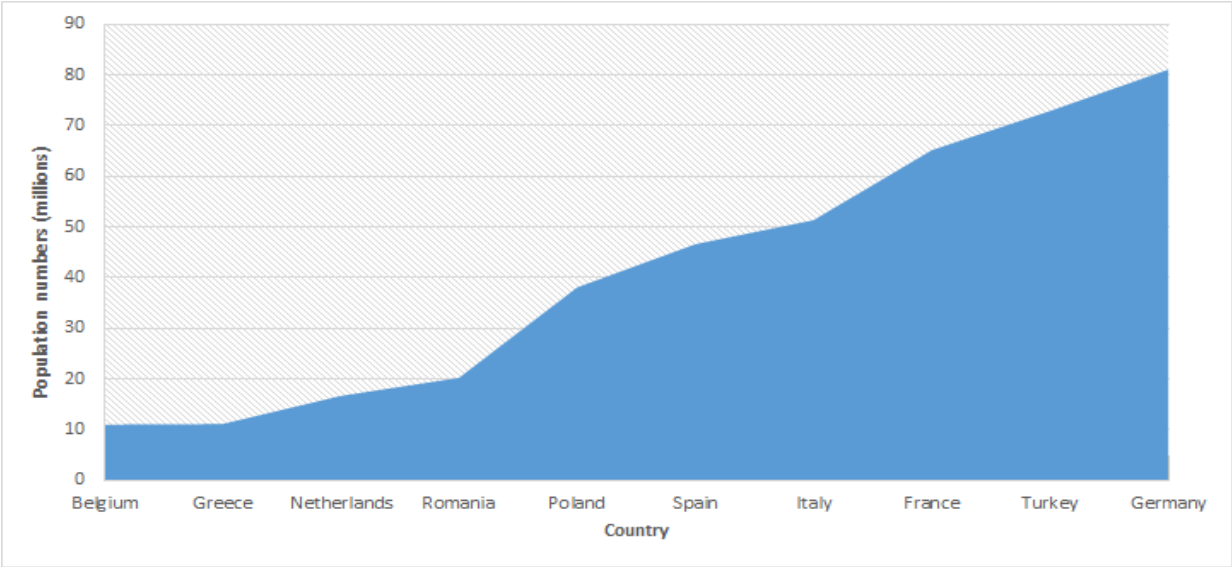


Figura 20: "Area plot" com *top* 10 dos países de origem dos estudantes do programa Erasmus em termos de população.

Na Figura 21 é possível observar o rácio entre o número de estudantes em Erasmus e a respectiva população dos países. Pode-se concluir que os países com maior população mais facilmente conseguem um rácio maior por milhão de habitantes.

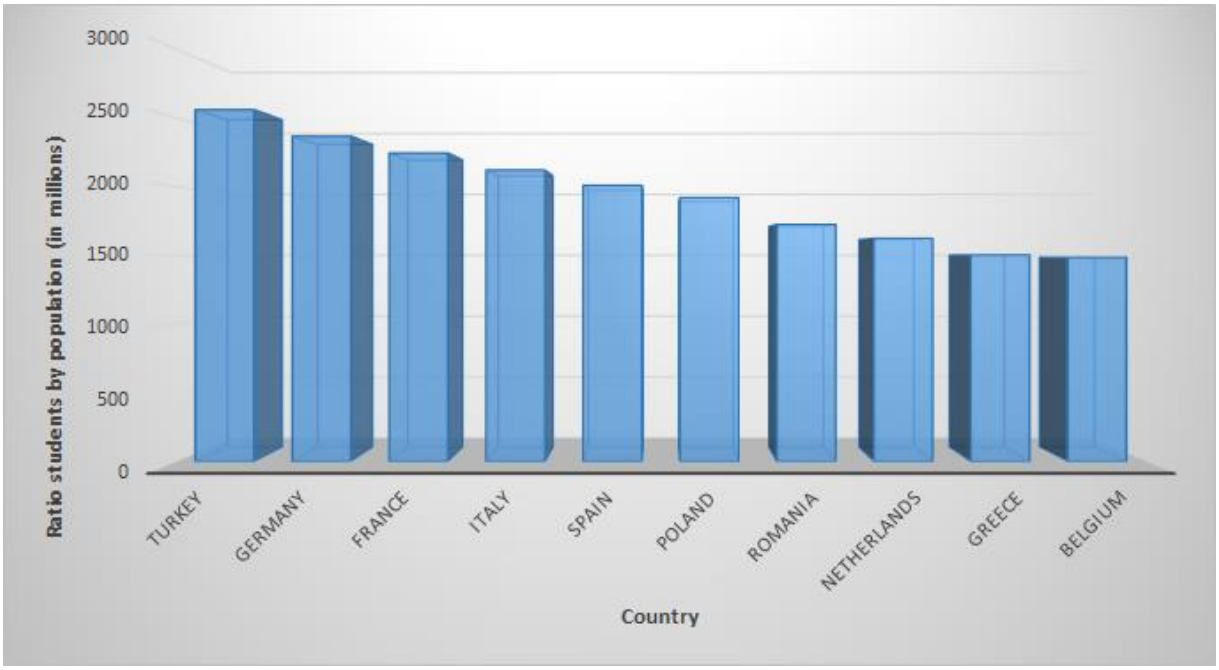


Figura 21: “Column plot” 10 países de origem dos estudantes do programa Erasmus em função da população (por milhão).

15.3. Quais as áreas de estudo em que foram efectuados mais ECTS?

Através da elaboração de um *stacked barplot* (Figura 22) utilizando a ferramenta Power BI Desktop, podemos observar que o *top* 10 das áreas de estudo com o maior número de ECTS, nos últimos 4 anos, foi realizado nas áreas de Línguas estrangeiras, Negócio e Administração, Humanidades, outro tipo de estudo não especificado, Direito, Economia, Ciência Política, Engenharia, Arquitectura, Medicina e Língua Nativa. Em geral, este padrão mantém-se para cada ano lectivo, com excepção da área das Humanidades, que apresentou menor número total de ECTS no ano lectivo 2011/2012, relativamente ao ano lectivo 2009/2010.

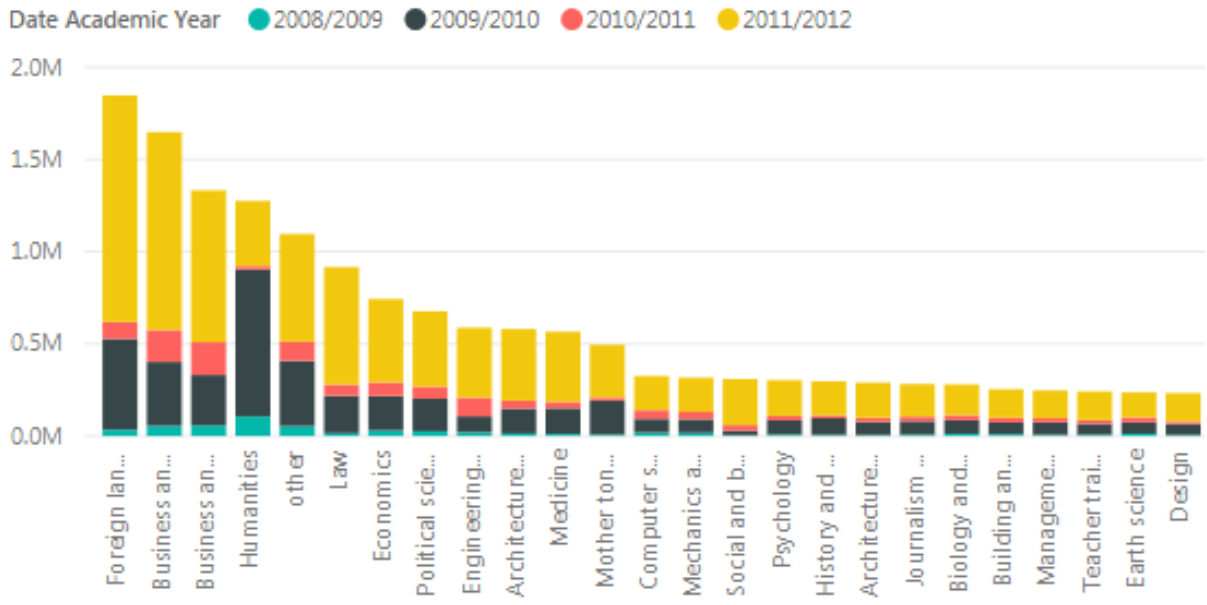


Figura 22: 'Stacked barplot' com o *top* das áreas de estudo em que foram efectuados mais ECTS.

15.4. Como varia o número de participantes de acordo com o género?

De acordo com o gráfico da Figura 23, observa-se um maior número de participantes do género feminino, comparativamente ao género masculino, respectivamente, cerca de 350 mil participantes eram do género feminino, e cerca de 220 mil participantes eram do género masculino. A evolução ao longo dos anos lectivos (Figura 24) mostra igualmente o mesmo padrão.

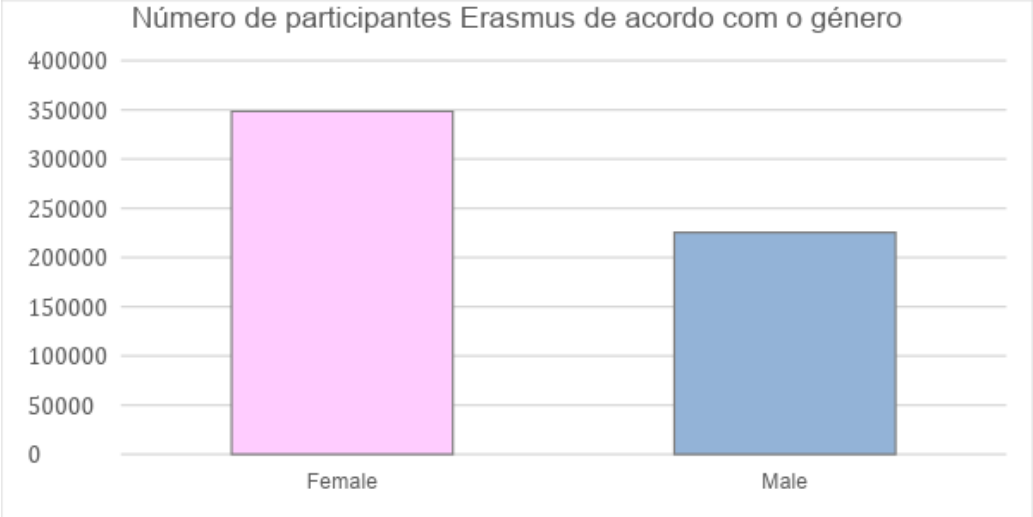


Figura 23: Número de participantes Erasmus de acordo com o género.

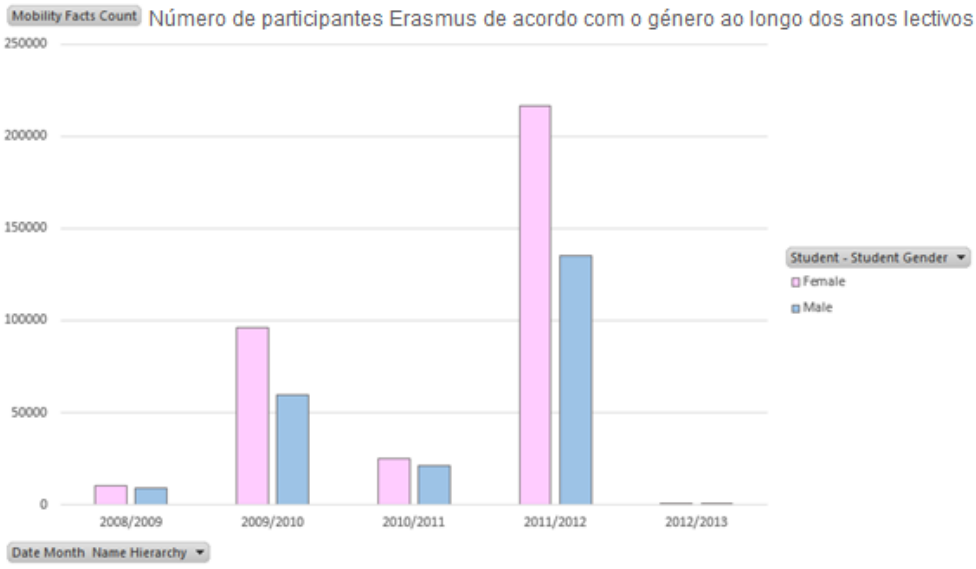


Figura 24: Evolução do número de participantes Erasmus de acordo com o género.

Obtendo os participantes de acordo com o género e país de origem (Figura 25), verifica-se que Portugal é o país com maior equilíbrio de participantes femininos e masculinos, tendo cerca de 7 mil participantes de cada um dos géneros feminino e masculino. Em geral, nos restantes países observa-se sempre um maior número de participantes do género feminino.

Obtendo os participantes de acordo com o género e país de destino (Figura 26), verifica-se que Portugal recebeu mais participantes do género feminino. Apenas alguns países receberam mais participantes do género masculino, nomeadamente, Suécia, Polónia, República Checa, Lituânia e Roménia.

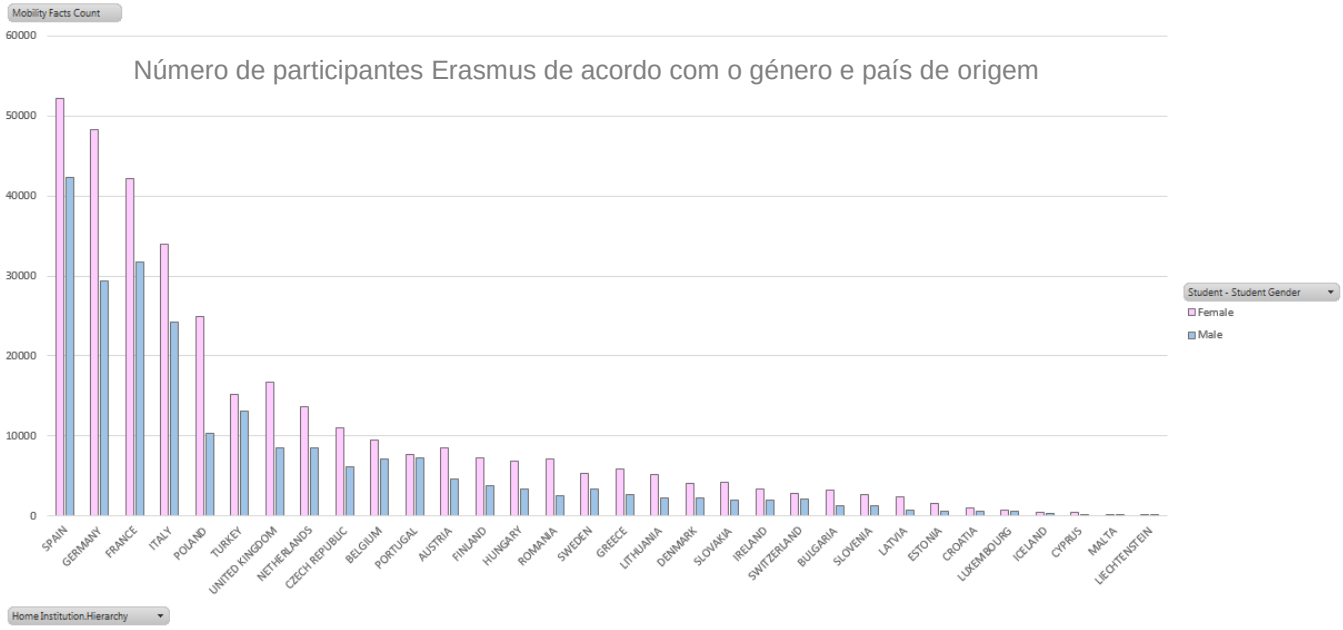


Figura 25: Número de participantes Erasmus de acordo com o género e país de origem.

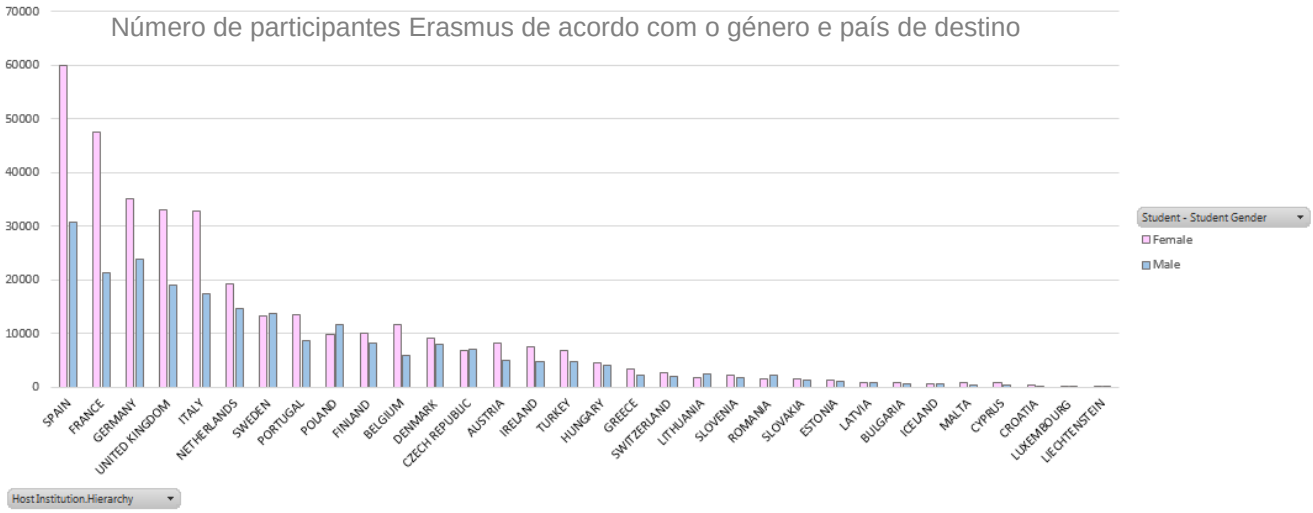


Figura 26: Número de participantes Erasmus de acordo com o género e país de destino.

Na figura 27 observa-se uma análise adicional do número de participantes de acordo com género e área de estudo. O *barplot* foi elaborado com as percentagens de participantes de acordo com o número de participantes de cada género – por exemplo, cerca de 13% dos participantes do género feminino tinham como área de estudo Línguas Estrangeiras (*Foreign Languages*), enquanto apenas cerca de 4% dos participantes do género masculino tinham esta área de estudo. Embora nos últimos anos tenham sido implementadas medidas, por parte do Instituto Europeu da Igualdade de Género, com o objectivo de eliminar desigualdades de género a nível académico e da investigação [7], o exemplo anterior, bem como outros, demonstram a existência de uma desigualdade de género relativamente às áreas de estudo escolhidas pelos participantes, no período de tempo referente aos dados analisados. Enquanto Línguas Estrangeiras teve um maior número de participantes femininos, Engenharia e Comércios de Engenharia (E*ngineering and Engineering Trades*) foi uma área selecionada maioritariamente por participantes do género masculino, bem como Ciência Computacional (*Computer Science*), e, Mecânica e Metalurgia (*Mechanics and Metal Work*). Houve ainda uma maior aderência às áreas de estudo relacionadas com Negócios, Administração e Economia por parte dos participantes do género masculino.

Por outro lado, a área de estudo de Língua Nativa (*Mother Tongue*) teve uma maior aderência por parte de participantes do género feminino, bem como Psicologia (*Psychology)*, Jornalismo e Informação (*Jornalism and Information*), e, Ciências da Educação (*Teacher Training and Education Science*).

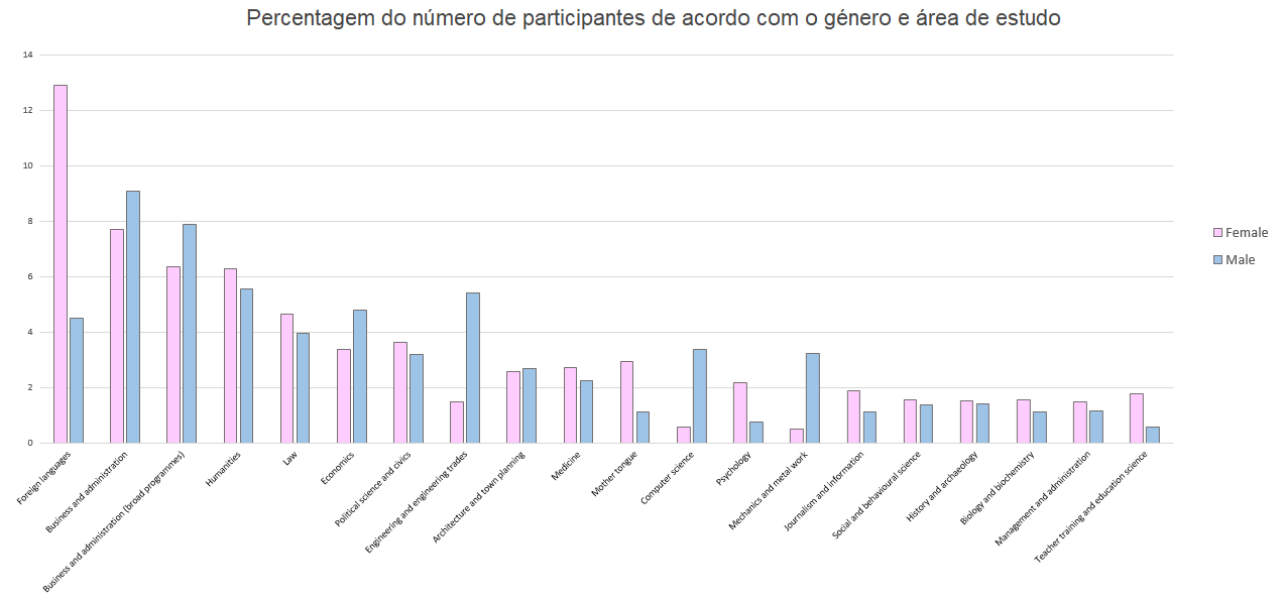


Figura 27: Número de participantes Erasmus de acordo com o género e área de estudo.

15.5. Qual o perfil de estudo dos estudantes portugueses?

O gráfico da Figura 28 foi elaborado utilizando um filtro de instituição hospedeira (*Home Institution*), no qual foi selecionado Portugal, obtendo-se assim o número de participantes portugueses de acordo com o país de destino.

Verifica-se que Espanha é o país de destino predilecto dos participantes de Erasmus portugueses, seguindo-se Itália e Polónia.

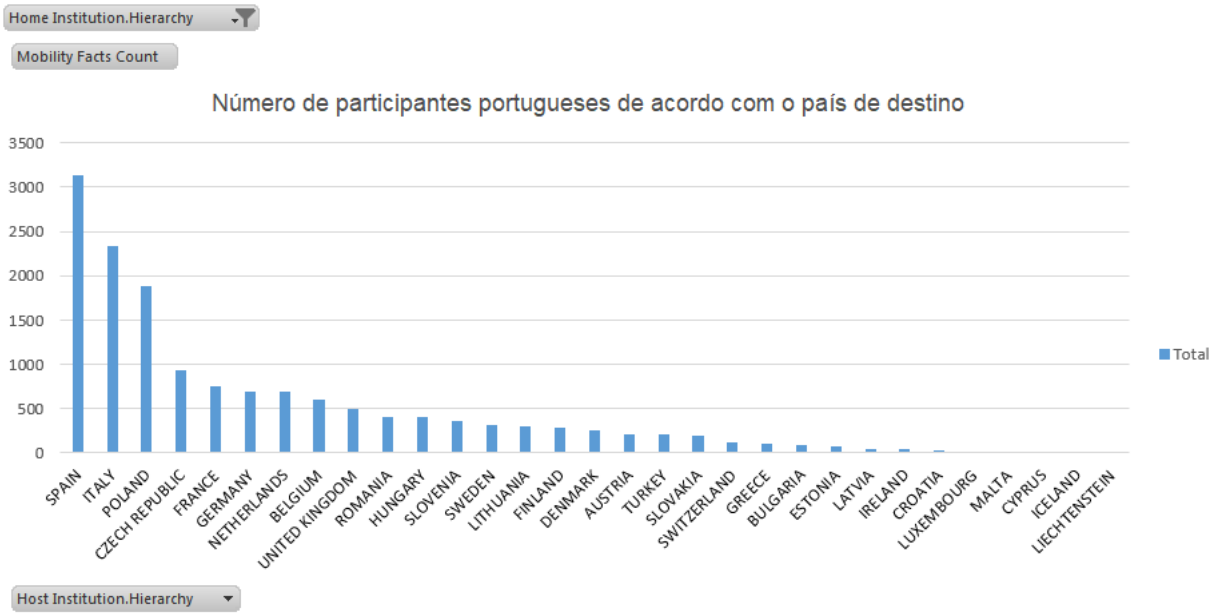


Figura 28: Número de participantes portugueses de acordo com o país de destino.

De seguida apresentam-se vários gráficos relativos ao perfil de estudo dos participantes portugueses. É possível concluir que as idades dos participantes se se incluíam maioritariamente entre os 19 e os 25 anos, com um número reduzido de participantes entre os 26 e 34 anos (Figura 29).

A maioria dos participantes pertencia ao primeiro ciclo (licenciatura), seguindo-se o segundo ciclo (mestrado), com apenas um número muito reduzido de participantes de cursos de curta duração e de terceiro ciclo (doutoramento) (fFgura 31). Relativamente à qualificação obtida no país de destino, a maioria dos participantes não iria obter qualquer qualificação durante o intercâmbio (Figura 32).

Relativamente à língua de aprendizagem, a maioria dos participantes não teve qualquer preparação/curso relativo à língua nativa do país de destino (Figura 32). Um número menor de participantes teve preparação da língua nativa do país de destino no próprio país de destino, seguindo-se uma preparação no país de origem, e cursos intensivos de Línguas Erasmus (EILC). Cerca de metade dos participantes não aprendeu na língua nativa do país, sendo que, nesses casos, a maioria teve aulas em inglês, como seria expectável, seguindo-se, em menor número, uma vasta gama de línguas, que incluem o Espanhol, Dinamarquês, Holandês, Francês, entre outras (Figura 32)

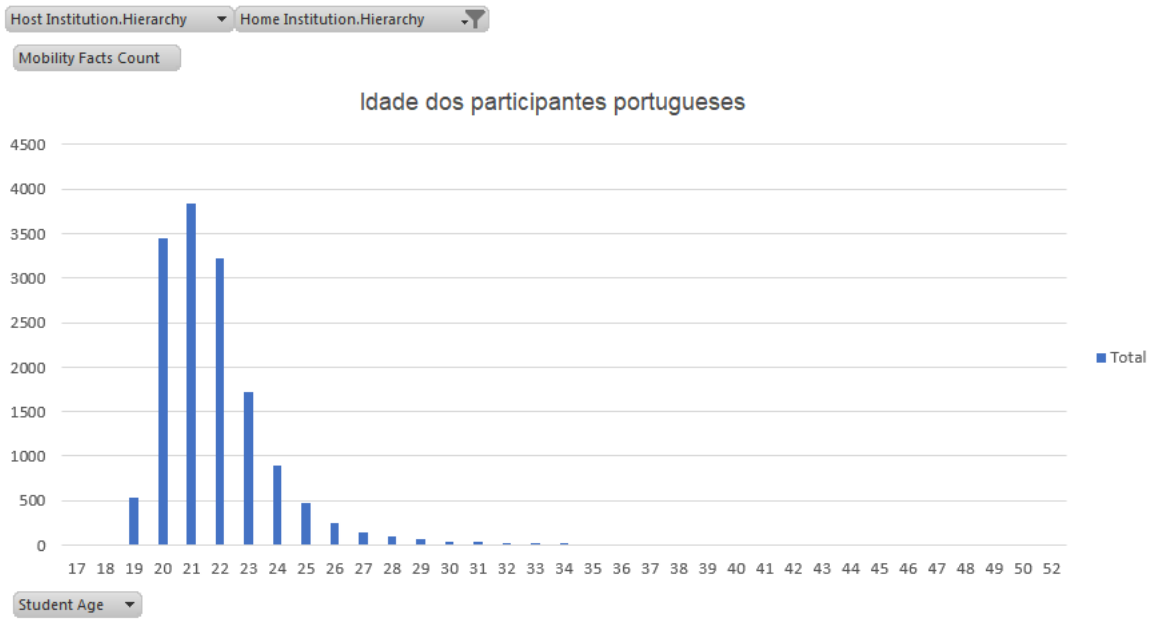


Figura 29: Idade dos participantes portugueses.

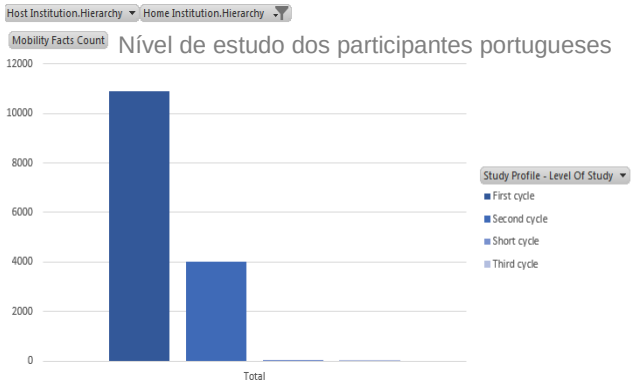


Figura 30: Nível de estudo dos participantes portugueses.

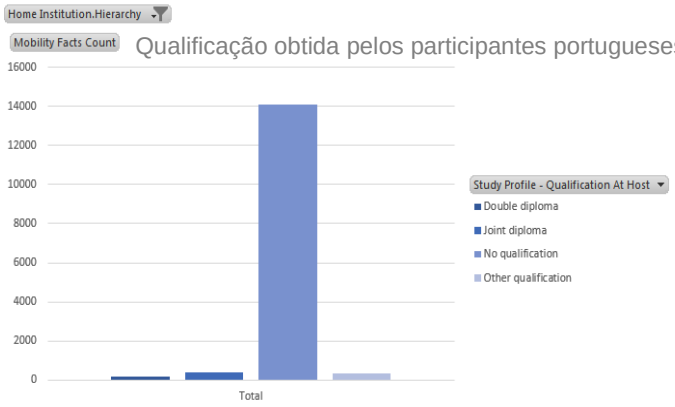


Figura 31: Qualificação obtida pelos participantes portugueses.

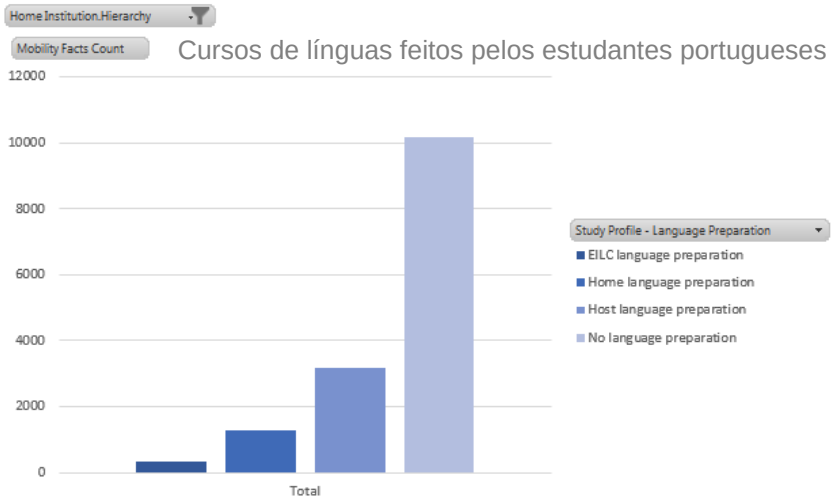


Figura 32: Cursos de línguas efectuados pelos estudantes portugueses, de forma a obter preparação para falar a língua do país de destino.

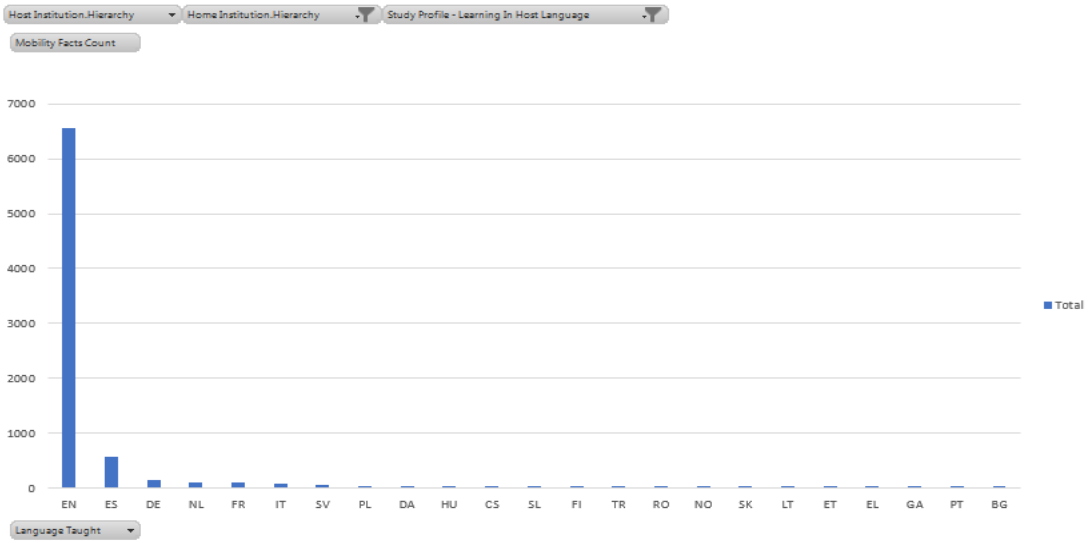


Figura 33: Línguas em que as aulas foram dadas, relativamente aos participantes portugueses que não aprenderam na língua nativa do país.

16. Métodos de prospecção de dados para o negócio

16.1. Agrupamento (*Clustering*)

Agrupar os países aderentes do programa Erasmus segundo o seu grau de semelhança, mediante as seguintes características dos estudantes: tempo de duração do programa, total de ECTS concluídos, valor da bolsa atribuída e número total de alunos por país inscrito no programa Erasmus. O critério de semelhança depende do algoritmo.

16.1.1. Preparação dos dados para o método

Foram usados, no método, 4 variáveis de estudo para agrupar os países (casos) presentes no programa Erasmus. As variáveis relativas aos estudantes de cada país são ‘Lenght of study period’, ‘Total ECTS’, ‘Study scholarship’ e soma total de alunos por país inscrito.

Na preparação dos dados para análise, uma vez que existem mais de meio milhão células de dados, foi calculada a média para cada variável, exceto na qual foi calculado o número total de alunos.

Para agrupar os dados baseados nas características idênticas foi utilizado a abordagem de agrupamento hierárquico aglomerativo, em que os *clusters* são constituídos por *subclusters.* Sendo um método não supervisionado, é desnecessário conhecer os agrupamentos ou *clusters* previamente à análise.

O método de agrupamento iterativo K-Means, onde o número de *clusters* ou k é escolhido aleatoriamente antes da análise, foi também testado mas apesar de os resultados terem sido coerentes apenas se conseguiu maus agrupamentos, tendo sido provavelmente má escolha de k, uma vez que o conteúdo dos *clusters* está muito dependente da escolha de k.

16.1.2. Agrupamento Hierárquico Aglomerativo

Foram usados 3 critérios de distância para juntar os *clusters*:

a. Single Linkage

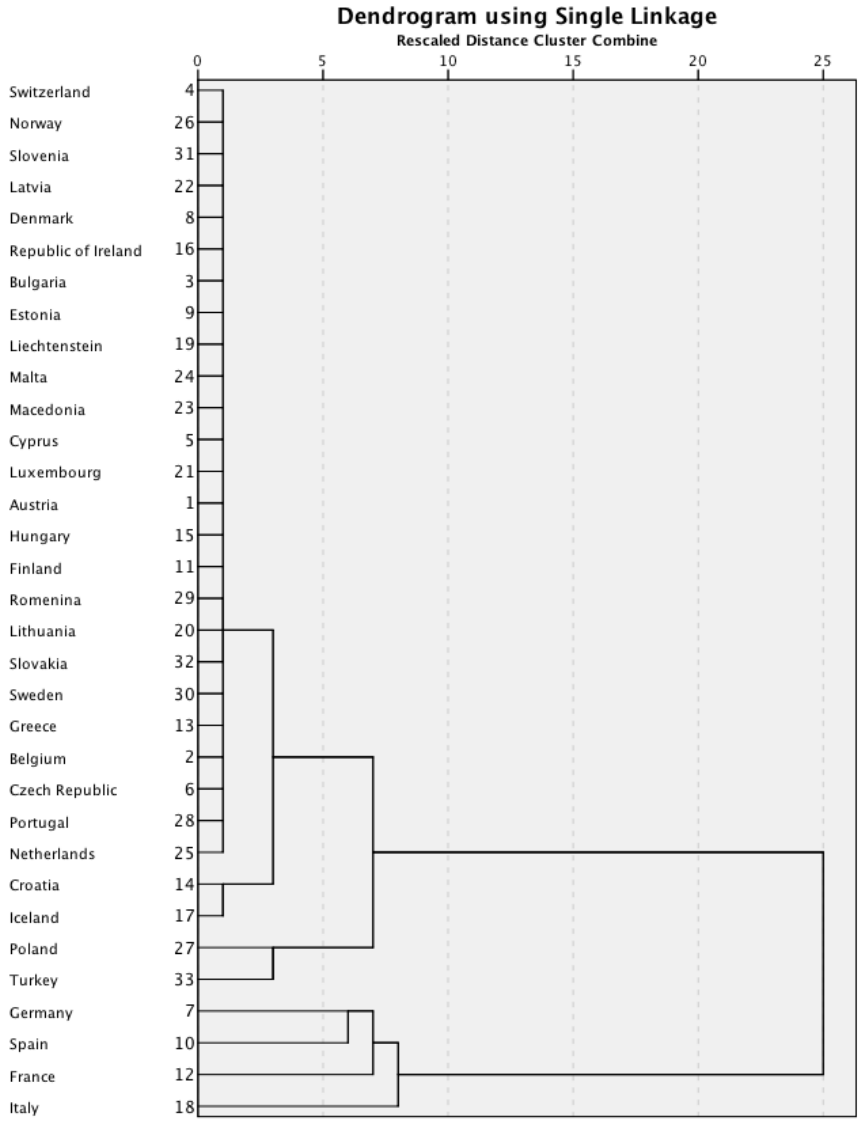


Figura 32: método hierárquico aglomerativo por single linkage (nearest neighbor)

Os *clusters* são agrupados de acordo com os pontos mais próximos entre si. Na análise é de salientar 2 grandes principais *clusters* no topo, que representam os *clusters* mais distantes e deste modo mais distintos entre si. Um dos *clusters* principais contém apenas *subclusters* com 4 pontos individuais, os países Alemanha, Espanha, France e Itália, enquanto que o outro *cluster* principal apresentam o mesmo número de *subclusters* mas com um maior número de pontos individuais, representantes dos outros

países no programa Erasmus.

b. Average Linkage

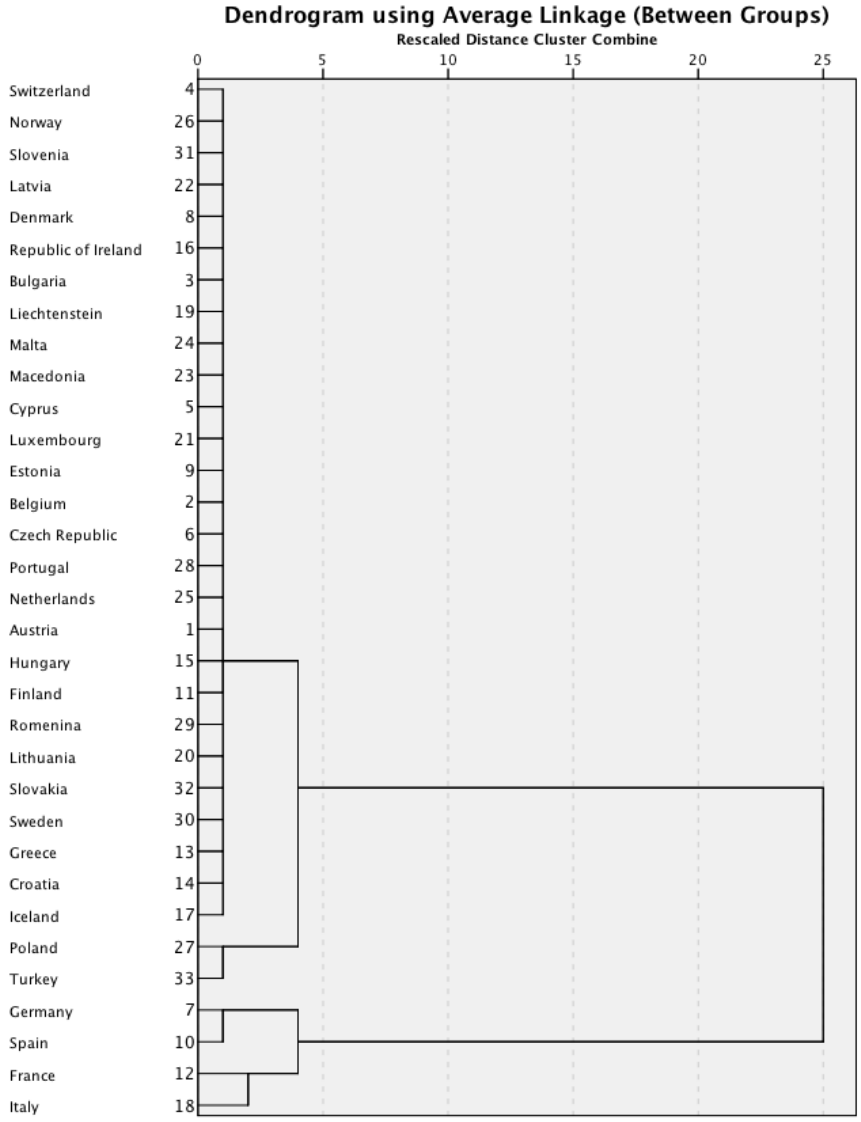


Figura 33: método hierárquico aglomerativo por average linkage (between groups)

Os *clusters* são agrupados de acordo com os pontos médios mais próximos entre si. Nesta análise é também de salientar os 2 grandes principais *clusters* no topo, que representam os *clusters* mais distintos entre si. Mais uma vez, os *clusters* referentes aos países Alemanha, Espanha, França e Itália estão muito próximos entre si, o que indica dados muito semelhantes. Existe um *subcluster* com os países Polónia e Turquia, e um outro com os restantes países.

c. Complete Linkage

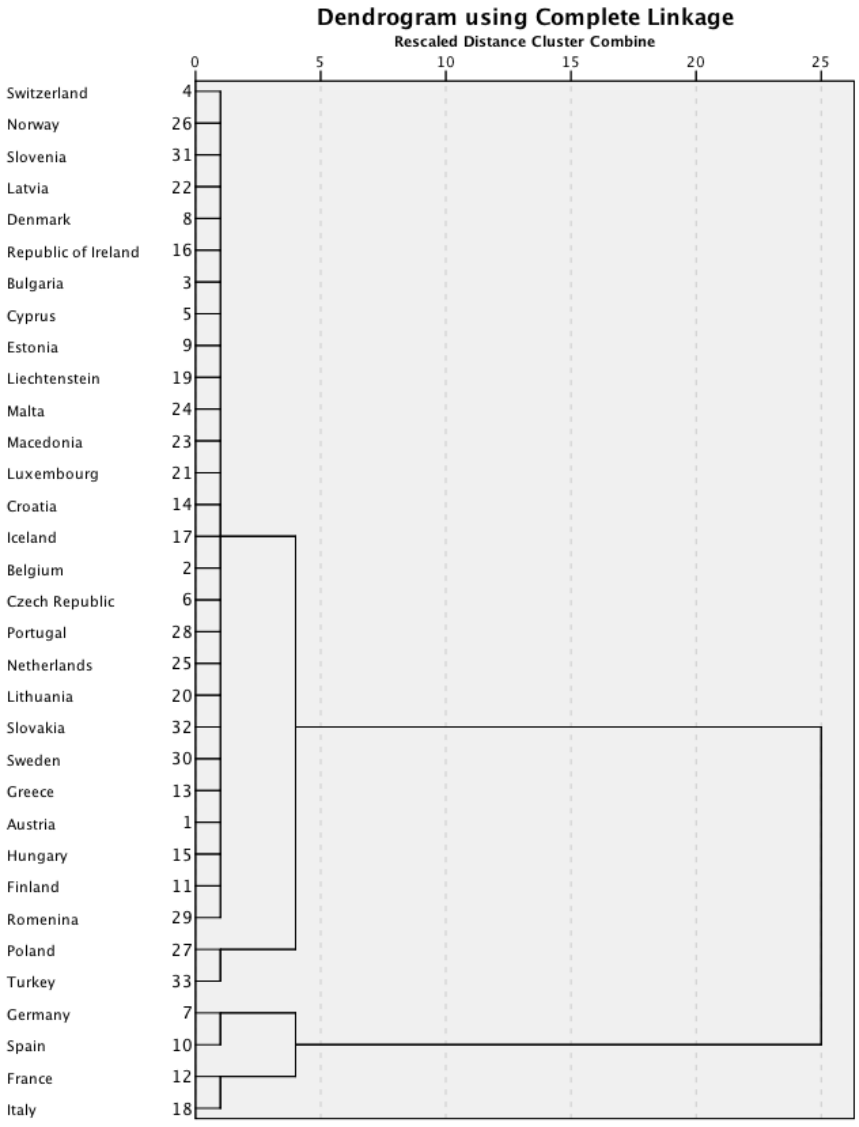


Figura 34: método hierárquico aglomerativo por complete linkage (furthest neighbor)

Os *clusters* são agrupados de acordo com os pontos mais distantes entre si. É novamente de salientar os 2 grandes principais *clusters* no topo, que representam os *clusters* mais distintos entre si. Mais uma vez, a presença de dois *subclusters* referentes aos países Alemanha, Espanha, França e Itália estão muito próximos entre si, mas desta vez a análise considera os países França e Itália ainda mais semelhantes do que anteriormente. Aqui também considerados estão os *subclusters* dos países Polónia e Turquia, e dos restantes países.

16.2. Método de Classificação com Redes Neuronais

O segundo método escolhido foi o método de classificação com redes neuronais. A ideia é criar uma rede que consegue prever o perfil de estudantes que fazem Erasmus em Portugal. Esta rede pode ser útil no planeamento de estratégias de forma a que os estudantes escolham Portugal como destino de Erasmus. Esta rede foi criada com o pacote neuralnet do R, o código encontra-se nos anexos enviados com o relatório. Usaram-se 3 conjuntos de dados de entrada, a idade(age), o sexo(gender) e a nacionalidade(nationality). Os dados previstos/saída foram Sim(Y), se CountryCodeOfHomeInstitution for PT e Não(N) caso contrário.

16.2.1. Preparação dos dados para o método

Este método é baseado em aprendizagem supervisionada e como tal a aprendizagem é feita à custa de exemplos pré-classificados. Na preparação dos dados foi necessário seleccionar os dados de interesse. Para o método ser aplicado não podem existir valores em falta(NA's), este problema foi resolvido na etapa anterior do projeto (Etapa III). Todos os dados texto foram transformados em unidades numéricas e foi feita a combinação dos dados de interesse. Não foi possível utilizar a totalidade dos dados (aproximadamente 500 mil linhas) pelo que se desenvolveu uma função que recolhesse várias amostras aleatórios. Destas amostras, 70% dos dados serviram como dados treino e os restantes 30% como dados teste de forma a treinar e testar a rede.

Passos para a preparação dos dados:

1ºPasso: Transformar os dados relativos à coluna CountryCodeOfHostInstitution, todos os dados "PT" subsitituiram-se por "Y" e todos os restantes por "N";

2ºPasso: Uma vez que os nossos dados têm aproximadamente 500 mil linhas, existe a probabilidade de ao gerar uma amostra aleatória de 1000, esta não inclua nenhum caso onde o estudante escolheu Portugal para fazer Erasmus e, por essa razão separaram-se os dados em 2 subconjuntos, dados onde o país de acolhimento é Portugal e outro conjunto de dados com os restantes paises.

3ºPasso: Criou-se uma função (randomSample) que retira n linhas de uma tabela. Esta função foi útil para retirar x amostras do subconjunto de dados relativos a Portugal e para retirar x amostras do outro subconjunto.

4ºPasso: Transformaram-se as variáveis categóricas, Sexo e Nacionalidade em factores. Depois foram convertidos em números uniformemente dispersos entre 0 e 1. Também se normalizou a variável contínua Idade.

5ºPasso: Converteu-se CountryCodeOfHostInstitution a uma estrutura semelhante a um índice de mapas de bits(com colunas Y e N) apropriadas para a rede neuronal.

16.2.2. Avaliação dos resultados- Método de Classificação com Redes Neuronais

Apesar do algoritmo ter convergido não podemos confiar no modelo, uma vez que tem um erro associado muito elevado (ver modelos nos anexo 4, 5 e 6). A grande taxa de erro sugere que o modelo não faz um bom trabalho ao prever o resultado nos dados teste. Podem existir três causas possíveis, primeiro, existe a possibilidade de overfitting. A nossa rede neuronal é sensível a esse problema, porque estamos a permitir um grande número de nós intermédios em relação aos dados de entrada. Portanto, a rede pode fazer um óptimo trabalho ao prever o resultado na amostra de treino. No entanto, a previsão muito personalizada para a amostra de treino, é mal executada na amostra de teste. Segundo, pode simplesmente não existir muita informação relevante para o nó de saída nos nós de entrada. Terceiro, o facto de apenas usarmos 1000 amostras pode não ser suficiente. Acreditamos que esta segunda explicação é mais plausível, porque encontramos erros semelhantes ao simplificar a rede neuronal, por exemplo, reduzindo o número de nós intermédios.

Podemos observar na tabela 14 que o erro decresce à medida que se acrescentam nós, no entanto mantem-se muito elevado. Em relação às métricas precisão e rechamada, mantiveram-se praticamente ao mesmo nível nos vários modelos. No entanto a rechamada é mais alta em todos, o que indica que estes modelos têm mais dificuldade em destacar os positivos. Relativamente ao resultado com maior probabilidade de ocorrência segundo os vários modelos, nenhum conseguiu prever com sucesso o resultado do teste. Por exemplo, no modelo com 4 nós intermédios, aplicando teste Portugal(Y) foram obtidas um maior número de colunas Portugal(Y), no entanto quando aplicado teste Outros(N) também foram obtidas um maior número de colunas Portugal(Y).

Tabela 15. Comparação dos resultados obtidos dos vários modelos de rede neuronal, com nós intermédios diferentes, que variam entre 4 a 7.

Camada intermédia= 4 nós		
Colunas teste	Colunas obtidas	
	Portugal(Y)	Outros(N)
Portugal(Y)	120	42
Outros(N)	101	37
Precisão: 0.543 Rechamada: 0.741 Erro: 171.224		
Camada intermédia= 5 nós		
Colunas teste	Colunas obtidas	
	Portugal(Y)	Outros(N)
Portugal(Y)	87	75
Outros(N)	91	47
Precisão: 0.489 Rechamada: 0.537 Erro: 164.635		
Camada intermédia= 6 nós		
Colunas teste	Colunas obtidas	
	Portugal(Y)	Outros(N)
Portugal(Y)	104	58
Outros(N)	97	41
Precisão: 0.517 Rechamada: 0.642 Erro: 163.889		
Camada intermédia= 7 nós		
Colunas teste	Colunas obtidas	
	Portugal(Y)	Outros(N)
Portugal(Y)	79	83
Outros(N)	105	33
Precisão: 0.429 Rechamada: 0.488 Erro: 150.615		

17. Conclusão

Este trabalho teve como objetivo a elaboração de um *Data Warehouse* de dados de mobilidade Erasmus. A concretização de cada etapa permitiu consolidar conhecimentos relativos às várias etapas necessárias para a elaboração deste sistema e, finalmente, transformar dados em informação útil no âmbito dos intercâmbios europeus.

Foram ainda encontradas algumas dificuldades, nomeadamente, o dicionário de códigos de instituições envolvidas nos dados de mobilidade não possuía os códigos necessários para englobar todas as instituições necessárias, pelo que, no futuro, seria necessário obter um dicionário de códigos atualizado.

A existência de anos académicos com um reduzido número de registos (2008/2009, 2012/2013) também foi um factor que dificultou a interpretação dos dados numa fase inicial. Deste modo, a obtenção de dados completos, caso possível, iria enriquecer as análises a efectuar no *Data Wharehouse* implementado.

O projecto *Integration services*, criado com o objectivo de importar os dados e criar automaticamente as tabelas relacionais na base de dados TPD14, foi uma tarefa bastante exigente ao nível do esforço e do tempo necessário para a executar. O principal motivo prende-se com a falta de familiaridade dos membros do grupo com as funcionalidades do *software Visual Studio* e com a dificuldade em detectar a origem dos erros que foram consecutivamente aparecendo no decorrer do projecto.

As dificuldades relatadas anteriormente foram ainda mais acentuadas na tarefa de criação e implementação do cubo através do mesmo software, cuja execução teve uma carga horária muito elevada.

17. Bibliografia

[1] www.scimagoir.com/rankings.php

[2] data.oecd.org/conversion/purchasing-power-parities-ppp.htm.

[3] data.worldbank.org/

[4] eacea.ec.europa.eu.

[5] kirste.userpage.fu-berlin.de/diverse/doc/ISO_3166.html

[6] www.umk.pl/en/erasmus/downloads/ISCED97_Erasmus_subject_codes.pdf

[7] eige.europa.eu/sites/default/files/mh0716096enn.pdf

18. Anexos

Anexo 1. Comparação dos 3 ficheiros iniciais: 'student_data_2009.csv', 'student_data_2010.csv' e 'student_1112.csv'.

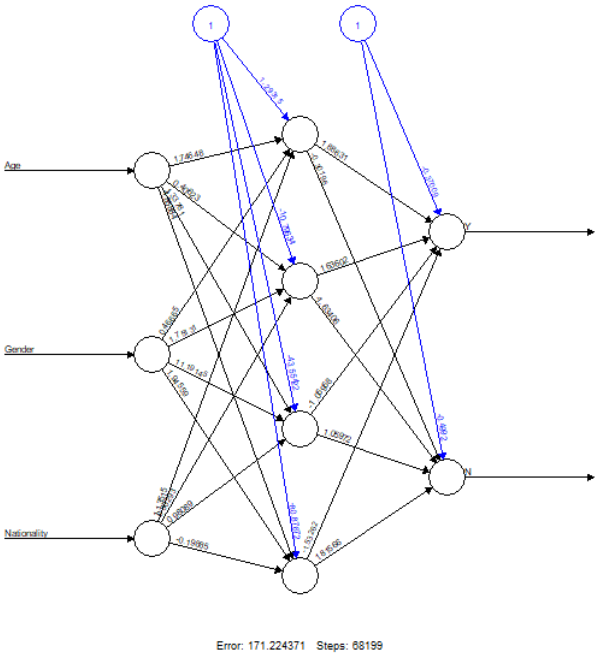
	student_data_2009.csv	student_data_2010.csv	student_1112.csv
NUMBER OF LINES BEFORE CLEANING:	213266	231408	252827
NUMBER OF LINES AFTER CLEANING:	176809	186676	200513
NUMBER OF LINES DELETED:	36457	44732	52314
NUMBER OF LINES WITH BLANKS:	35561	40909	48071
NUMBER OF LINES WITH XX:	894	3823	4243

Anexo 2. Comparação entre o ficheiro com os dados originais ("Student_mobility_raw.csv") e os dados após o processo de limpeza ("Student_mobility_clean.csv").

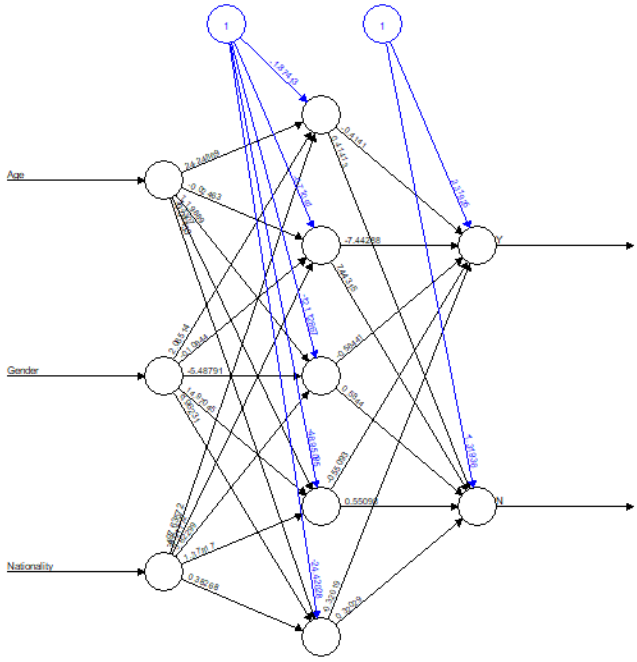
Ano Lectivo	Student_mobility_raw.csv	Student_mobility_clean.csv
ano_lec_8/9:	19629	19520
ano_lec_9/10:	158011	157185
ano_lec_10/11:	47992	47168
ano_lec_11/12:	361422	353754
ano_lec_12/13:	173	171
Ano		
8:	0	0
9:	127126	126454
10:	50597	50334
11:	293813	288045
12:	115691	112965
13:	0	0
Mês-ano		
jan-09:	7	7
fev-09:	16	16
mar-09:	0	0
abr-09:	1	1
mai-09:	0	0
jun-09:	114	113
jul-09:	623	617
ago-09:	18868	18766
set-09:	87244	86798
out-09:	19693	19583
nov-09:	452	445
dez-09:	108	108
jan-10:	15630	15548
fev-10:	25409	25279
mar-10:	5820	5782
abr-10:	3054	3043

mai-10:	297	297
jun-10:	229	227
jul-10:	66	66
ago-10:	9	9
set-10:	83	83
jun-11:	165	163
jul-11:	1013	997
ago-11:	46731	45925
set-11:	205525	201589
out-11:	39421	38453
nov-11:	771	741
dez-11:	187	177
jan-12:	41929	40945
fev-12:	55257	54107
mar-12:	11333	10983
abr-12:	5823	5643
mai-12:	499	479
jun-12:	473	439
jul-12:	141	137
ago-12:	63	61
set-12:	173	171

Anexo 3. Camada intermédia com 4 nós da rede neuronal.

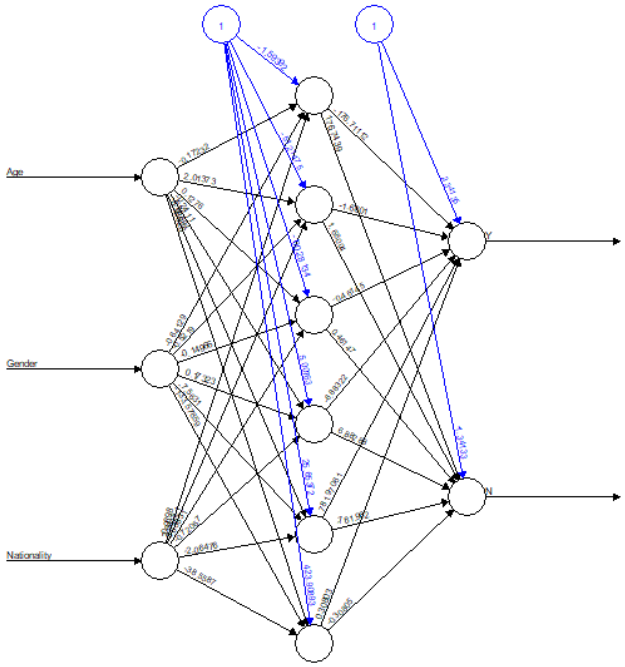


Anexo 4. Camada intermédia com 5 nós da rede neuronal.



Error: 164.634867 Steps: 108554

Anexo 5. Camada intermédia com 6 nós da rede neuronal



Error: 163.889236 Steps: 145098

Anexo 6. Camada intermédia com 7 nós da rede neuronal