# Information Visualization

# CHECKPOINT II: Data cleaning and processing

G21-A

**1. Initial Dataset**

We used data from four datasets:

1. **Student\_Mobility\_2013-2014.xlsx** - Original dataset contains 27 columns and 272 497 entries. Each record represents one student participating in the program.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ... | SendingCountry | ReceivingCountry | ... | LevelOfStudy | SendingPartnerErasmusID | HostingPartnerErasmusID | ... |
| ... | AT | NO | ... | Second Cycle | D OSNABRU01 | CZ PRAHA07 | ... |

1. **Cost\_of\_living\_2016.csv** - Contains data about cost of living in cities around the world. The dataset has 577 records and 21 coulmns. Each record represents one city.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| City | Country | Costofliving.index | ... | Domestic beer | One-way Ticket | Cinema | Apple |
| Aachen | Germany | 62.14 | ... | 1.00 | 2.93 | 8.93 | 2.05 |

1. **EUC\_for\_the\_academic\_year\_2013\_2014.xls** - Contains data about universities participating in the ERASMUS program. The dataset contains 4919 records and 7 columns.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Institutional code | Application Reference Number | Name of Organisation | Country | City |
| A BADEN01 | 28505-IC-1-2007-1-AT-ERASMUS-EUC-1 | PAEDAGOGISCHE HOCHSCHULE NIEDEROESTERREICH | AT | Baden |

1. **countryinfo.txt** - Contains data about the worlds countries. The dataset has 253 records and 19 columns.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ISO | ISO3 | ISO-Numeric | fips | Country | Capital | Area (in km2) | Population | Continent | tid | Currency code | .. |
| AD | AND | 20 | AN | Andorra | Andorra la Vella | 468 | 84000 | EU | .ad | EUR | .. |

**2. Selected/Derived Data**

* + - 1. **1) Student\_Mobility\_2013-2014.xlsx** - From this dataset, we selected *SendingCountry, ReceivingCountry, MobilityType, LevelOfStudy, ParticipantGender, SendingPartnerErasmusID* and *HostingPartnerErasmusID.* We also add counter for number of participants from sending country and from receiving country and also ratio of these two counters. We counted the amount of *Bachelors, Masters* and *PhDs* coming into the country using *LevelOfStudy*.
      2. **2) Cost\_of\_living\_2016.csv** - Grouped by *country*, and used average mean to aggregate *cost.of.living.index, rent.index* and *beer price.* Merge joined with table containing erasmus countries. Cost of living attribute was calculated per-country since joining based on a city name was found to be extremely challenging (duplicate cities, encoding errors, localized names, human-introduced typing errors).
      3. **3) EUC\_for\_the\_academic\_year\_2013\_2014.xls** - *Institutional code, Name of Organisation, Country* and *City* was used to obtain the coordinates of the participating institutions (we queried a Google Cloud Geolocation API).
      4. **4) countryinfo.txt** - We used *ISO* and *Country* to mapped ISO shortcuts from other datasets on country names.
      5. **3. Data abstraction**
      6. The datasets we use are all static tables. We also have geometry (spatial) dataset type for the position coordinates of the universities.

|  |  |  |
| --- | --- | --- |
| **column name** | **attribute description** | **semantics** |
| CountrySending, CountryReceiving, ISO | nominal | ISO country code |
| StudentIncomingGender | nominal | Gender of participating student |
| StudentIncomingCount, StudentOutgoingCount | ratio | Number of outgoing/incoming students for the country |
| In/Out | ratio | Ration between incoming and outgoing students |
| StudentBachelorIncomingCount, StudentMasterIncomingCount, StudentPhDIncomingCount | ratio | Number of participants studying specific degree |
| Country | nominal | Name of the country |
| Cost.of.Living.Index | continuous | Number that describes cost of living |
| Rent.Index | continuous | Number that describes cost of rent |
| Domestic Beer (0.5 liter bottle) | ratio | Price per domestic beer |
| InstitutionSending, InstitutionReceiving | nominal | Name of the university |
| CoordinatesSending, CoordinatesReceiving | continuous | Coordinates of sending / receiving university |

* + - 1. **4. Dataset processing**
      2. The dataset does contain missing values especially in the *HostingPartnerErasmusID*as in case of work exchange, the companies do not have an unique identifier. Since we are only concentrating on the student exchanges, we filtered the data out. The dataset does not contain any duplicates, however one big problem is with the broken encoding of the city names. In order to obtain readable city names, we had to join the original dataset with the EUC dataset. But even then the cities were written in different languages. So instead of using cities for positions, we obtain the coordinates of the participating institutions (we queried a Google Cloud Geolocation API).
      3. We joined the datasets with using country ISO codes and universities Erasmus codes.
      4. In the end we have eight datasets that we can use for answering our questions: *favourite\_20191018.csv* (2 columns, 33 rows), *gender\_20191018.csv* (3 columns, 66 rows), *favourit\_in\_out\_20191018.csv* (2 columns, 33 rows), *flow\_coordinates\_20191018.csv*   
         (7 columns, 77 326 rows), *geocodedMobility.csv* (5 columns, 207790 rows), *favourite\_degree\_20191018.csv* (4 columns, 32 rows), *flow\_20191018.csv* (3 columns,   
         928 rows), *cleaned-cost-of-living.csv* (4 columns, 34 rows).
      5. **5. Mapping (Data sample / Questions)**

|  |  |
| --- | --- |
| **CountryReceiving** | **StudentIncomingCount** |
| ES | 30 275 |
| FR | 24 057 |

|  |  |
| --- | --- |
| **CountrySending** | **In/Out** |
| MT | 3,178807947 |
| NO | 2,7124518614 |

|  |  |  |  |
| --- | --- | --- | --- |
| **CountrySending** | **StudentsBachelorIncomingCount** | **StudentsMasterIncomingCount** | **StudentPhdIncomingCount** |
| AT | 3 195 | 1 739 | 49 |
| BE | 4 477 | 1 925 | 45 |

1) Which countries are popular for being an Erasmus destination? 2) What is the ratio between incoming and

outcoming students in Norway?

|  |  |  |  |
| --- | --- | --- | --- |
| **ISO** | **Cost.of.Living.Index** | **Rent.Index** | **Domestic Beer (0.5 liter bottle)** |
| AT | 70.0 | 26.7 | 1.1766666667 |
| BE | 79.0 | 28.0 | 1.25 |

3)Are the most popular countries for bachelors different than for masters and PhD? 4) What are the differences in cost of living

between receiving countries?