Data Cleaning on dataset of postcards:

* Separated several columns (not limited to the geographical related ones) that can make it easier for further manipulation
* Original columns were kept for check

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No | Column Name | Description | Cleaning Process | Try out | Remark |
|  | Uniform title | Indicate the geographical location and main elements of the postcard image  Different elements separated by “. ” | 1. Separated columns by single value with new column names “Image Element (City)”, “Image Element 2”, “Image Element 3”, etc.  2. Further cleaned and clustered the “Image Element (City)” column as “Image Element (City)\_clean” | 1. Tried out a heat map on Tableau  2. Generated a word cloud for “Image Element 2” column on Tableau | 1. Vast majority of the values in column “Image Element (City)\_clean” could be automatically mapped to geographical coordinates, could be used by other methods  2. The separated elements may give a hint to image analysis? |
|  | Date | Indicate the date of publication of the postcard  Around 1/3 rows have no value | Extracted only numeric values in a new column “Date (numeric)”, and generalized the time period every 20 years in another column “Publication Period” (not accurate) | 1. Applied the period division as filter to different charts on Tableau | The period division is not accurate, but may serve as a reference |
|  | Colour | Indicate the colour type of the postcard image | Cleaned some noise and re-clustered the types in a new column “Colour (re-cluster)” | Tried out a bubble chart of distribution of different colour types in the dataset | When interacting with publication period, it shows a leading trend of postcards in colour and in black and white during different time periods. Maybe interesting to link with the publishing history (inks, materials, circulation, etc.) during two world wars? |
|  | Group title | Indicate the location of collection group in three languages (Dutch, English, French)  Texts in different languages normally separated by “;” | 1. Separated columns in single language with new column names as “Region (Dutch)”, “Region (English)”, “Region (French)”  2. Further cleaned and re-clustered the region names in English in a new column named “Region (Province name in English)”, only keep the Belgium province name for automatically mapping in Tableau | Tried out generating a heatmap with the quantity of postcards from each region on Tableau | The cleaned region information can be made use of by other methods |
|  | Language | Indicate the language of contents in the postcards | Cleaned and re-clustered the language codes | Tried out a line graph showing different language distribution overtime with publication year information | Found a fun fact: there are two postcards in the collection were in Esperanto, but we don’t have the time information about them |
|  | Label (Library Call number) | Indicate the affiliated library of the specific postcard collection | Extracted the abbreviation of libraries | N/A | Might be helpful if any exploration on library collections |

Some new ideas:

* Except for listed ones, “Main title” column (textual description about graphical content / theme of the postcard) could be further cleaned for text analysis and help with the goal of “cluster postcards topically”? (but a bit time consuming, since the multilingual values were not separated in united format)
* Might be able to do a network analysis taking “publisher”, “place of publication”, “region” as nodes