R Practical

Data importing, cleaning and basic descriptives

Xanthi Andrianou

ECDC fellowship Introductory Course

Spetses, October 2023

## Background

The practical is based on a scenario that requires the **analysis of surveillance data** coming from two sources: **cases-based** information reported by countries and **aggregated** case data collected from open sources. The datasets have been created to resemble the data collected during monitoring of mpox in Europe in 2022.

## Scenario

E-pox has been reported in 5 countries for the first time in 2022. We have collected aggregated case numbers from open sources, and we have also reported case-based data. The objectives of the practical are the following:

1. Explore different types of files and how they can be imported in R
2. Perform basic data cleaning, e.g., changing the variable type, recode variables, aggregate and filter
3. Perform a basic descriptive analysis using tables and graphs

## Scripts and output

|  |  |  |
| --- | --- | --- |
| Script | Output | Notes |
| E\_pox\_Rpractical\_basic.R | E\_pox\_Rpractical\_basic.docx | Script detailing all the analysis and the output (including code snippets and the result) |
| E\_pox\_Rpractical.Rmd | E\_pox\_Rpractical.html | R markdown script generating the output in html format |
| Install\_packages.R |  | Script to install the necessary packages (included also in the standalone scripts). It includes instructions to use renv (for reproducible R environment) |
| E\_pox\_Rpractical\_renv.R |  | Same as the E\_pox\_Rpractical\_basic.R but using the renv approach (renv is not activated for this project by default) |

## Codebooks

### Case-based data

|  |  |
| --- | --- |
| **Dataset name:** | E\_pox\_case\_based\_data |
| **Column count:** | 14 |
| **Row count:** | 2,000 |

**Column Attributes:**

|  |  |  |
| --- | --- | --- |
| **1** | **Column name:** | **DateOfNotification** |

|  |  |  |
| --- | --- | --- |
| **2** | **Column name:** | **DateOfDiagnosis** |

|  |  |  |
| --- | --- | --- |
| **3** | **Column name:** | **DateOfOnset** |

|  |  |  |
| --- | --- | --- |
| **4** | **Column name:** | **Age** |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **5** | **Column name:** | | **Gender** | | | |
| Categories | | Frequency | | Cumulative Frequency | Percent |
| F – Female | | 36 | | 36 | 1.80 |
| M – Male | | 1,960 | | 1,996 | 98.00 |
| O – Other (e.g., transgender man, transgender woman and collected as free text) | | 1 | | 1,997 | 0.05 |
| UNK - Unknown | | 3 | | 2,000 | 0.15 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **6** | **Column name:** | | **SexualOrientation** | | | |
| Categories | | Frequency | | Cumulative Frequency | Percent |
| BISEXUAL – Bisexual | | 7 | | 7 | 0.35 |
| HETERO – Heterosexual | | 46 | | 53 | 2.30 |
| MSM – Men who have sex with men/homo or bisexual male | | 833 | | 886 | 41.65 |
| UNK – Unknown or undetermined | | 214 | | 1,100 | 10.70 |
| Missing | | 900 | | 2,000 | 45.00 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **7** | **Column name:** | | **Outcome** | | | |
| Categories | | Frequency | | Cumulative Frequency | Percent |
| A – Alive | | 1,405 | | 1,405 | 70.25 |
| UNK – Unknown | | 595 | | 2,000 | 29.75 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **8** | **Column name:** | | **HIVStatus** | | | |
| Categories | | Frequency | | Cumulative Frequency | Percent |
| NEG – Negative | | 525 | | 525 | 26.25 |
| POS – Positive | | 307 | | 832 | 15.35 |
| UNK – Unknown | | 446 | | 1,278 | 22.30 |
| Missing | | 722 | | 2,000 | 36.10 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **9** | **Column name:** | | **Country** | | | |
| Categories | | Frequency | | Cumulative Frequency | Percent |
| CountryA | | 816 | | 816 | 40.80 |
| CountryB | | 391 | | 1,207 | 19.55 |
| CountryC | | 474 | | 1,681 | 23.70 |
| CountryD | | 217 | | 1,898 | 10.85 |
| CountryE | | 102 | | 2,000 | 5.10 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **10** | **Column name:** | | **ClinicalSymptoms** | | | |
| Categories | | Frequency | | Cumulative Frequency | Percent |
|  | | 12 | | 12 | 0.60 |
| Lesions | | 14 | | 26 | 0.70 |
| Rash | | 257 | | 283 | 12.85 |
| Rash, Lesions | | 323 | | 606 | 16.15 |
| Rash, Systemic symptoms | | 676 | | 1,282 | 33.80 |
| Rash, Systemic symptoms, Lesions | | 654 | | 1,936 | 32.70 |
| Systemic symptoms | | 28 | | 1,964 | 1.40 |
| Systemic symptoms, Lesions | | 36 | | 2,000 | 1.80 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **11** | **Column name:** | | **Rash** | | | |
| Categories | | Frequency | | Cumulative Frequency | Percent |
| Rash | | 1,910 | | 1,910 | 95.50 |
| Missing | | 90 | | 2,000 | 4.50 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **12** | **Column name:** | | **SystemicSymptoms** | | | |
| Categories | | Frequency | | Cumulative Frequency | Percent |
| Systemic symptoms | | 1,394 | | 1,394 | 69.70 |
| Missing | | 606 | | 2,000 | 30.30 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **13** | **Column name:** | | **Lesions** | | | |
| Categories | | Frequency | | Cumulative Frequency | Percent |
| Lesions | | 1,027 | | 1,027 | 51.35 |
| Missing | | 973 | | 2,000 | 48.65 |

### Aggregated data

|  |  |
| --- | --- |
| **Dataset name:** | E\_pox\_aggregated\_data |
| **Column count:** | 3 |
| **Row count:** | 101 |

**Column Attributes:**

|  |  |  |
| --- | --- | --- |
| **1** | **Column name:** | **DateRep** |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **2** | **Column name:** | | **Country** | | | |
| Categories | | Frequency | | Cumulative Frequency | Percent |
| CountryA | | 28 | | 28 | 27.72 |
| CountryB | | 26 | | 54 | 25.74 |
| CountryD | | 26 | | 80 | 25.74 |
| CountryE | | 21 | | 101 | 20.79 |

|  |  |  |
| --- | --- | --- |
| **3** | **Column name:** | **Cases** |