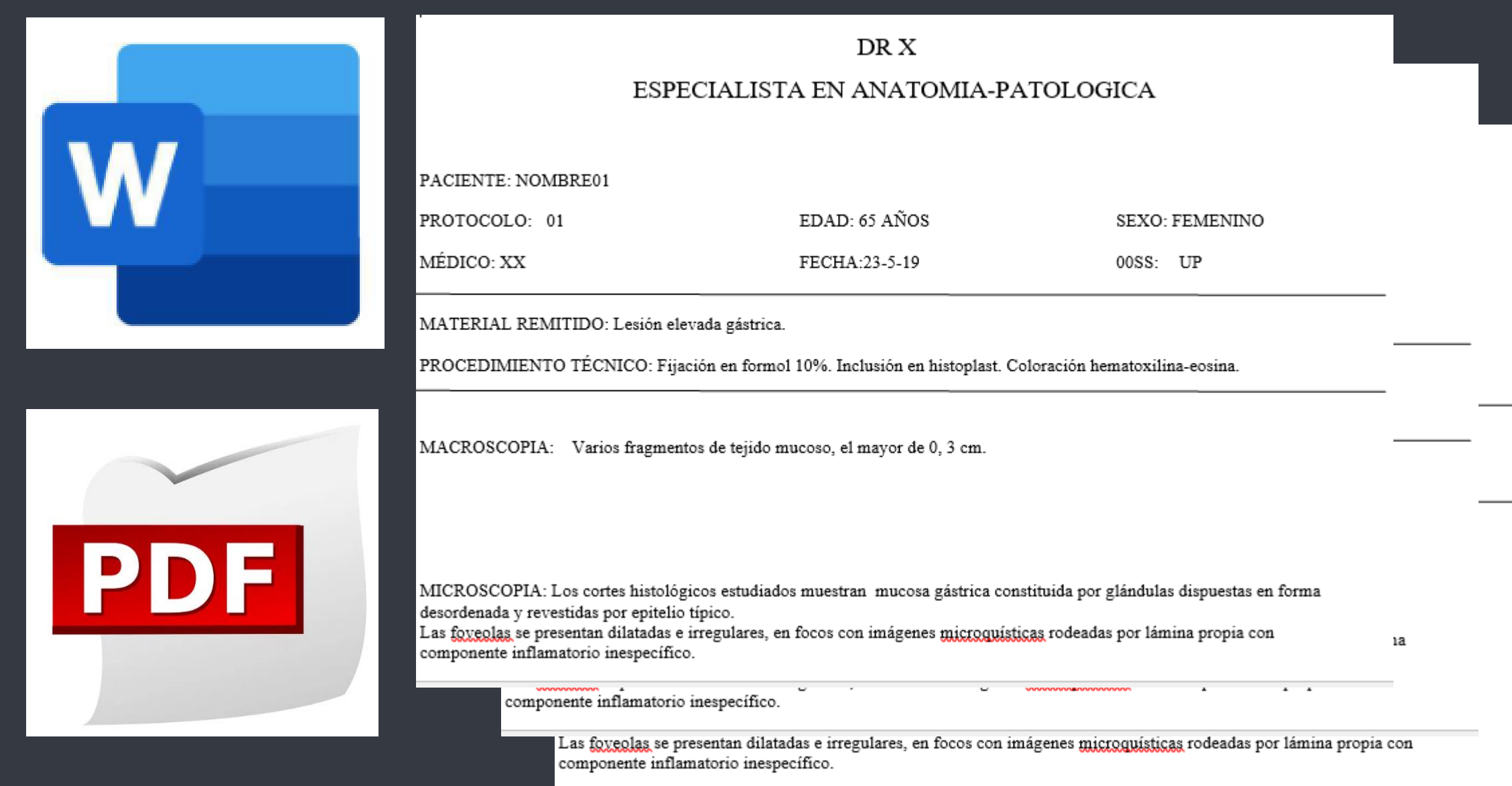
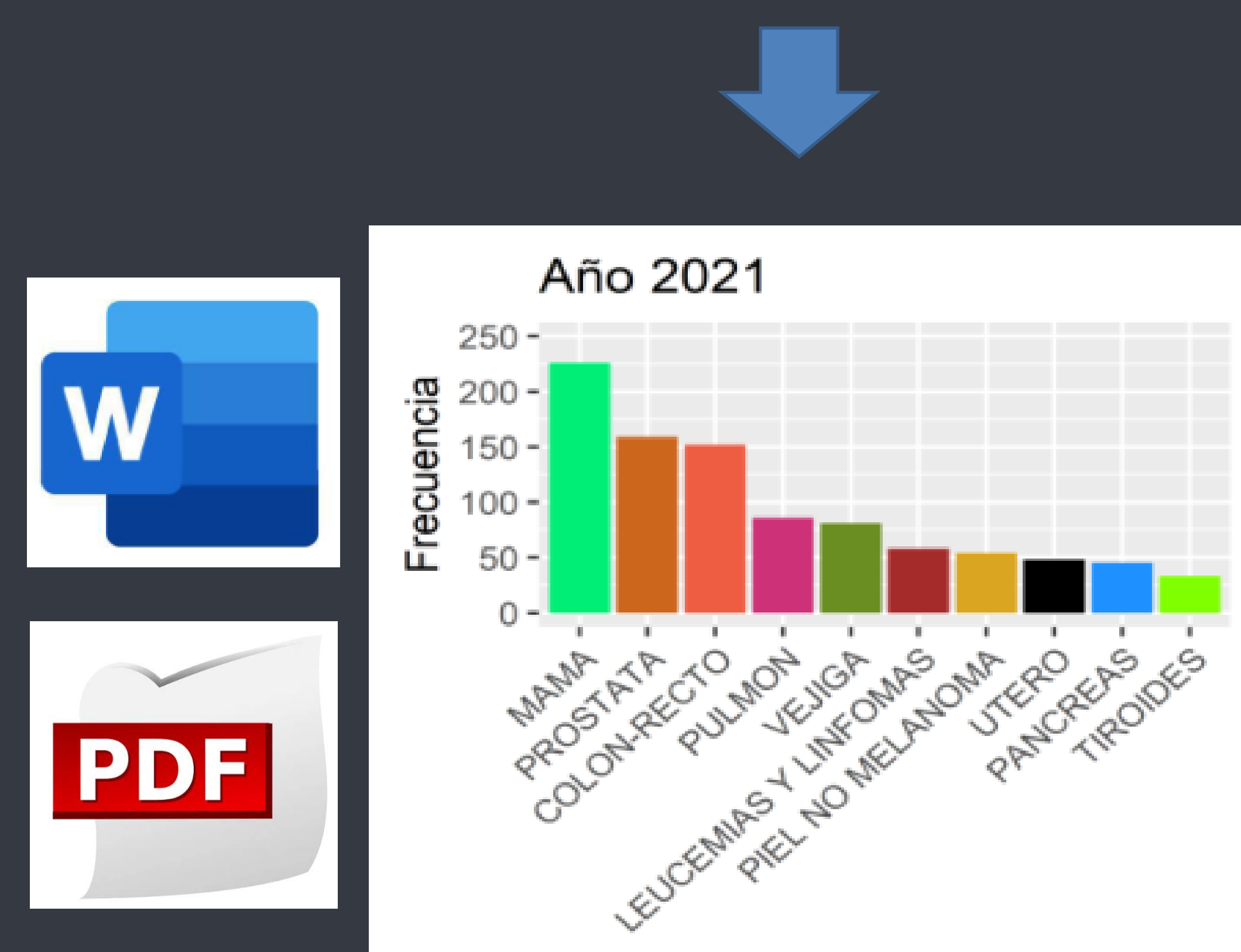


QUICK CANCER REPORT STARTING FROM A RAW DATABASE

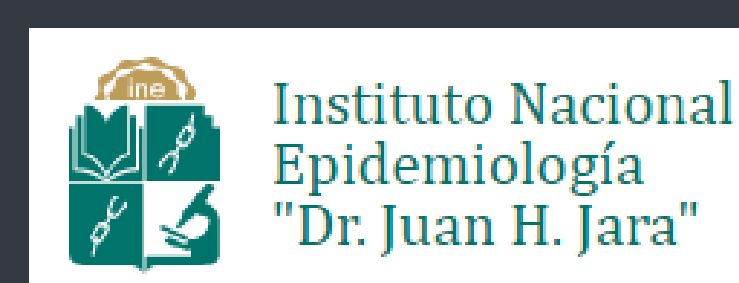


| | B | C | D |
|----|--|---|------|
| | A RECEP DIAGNOSTICO | | AÑO |
| | /4/2018 ADENOCARCINOMA MODERADAMENTE DIFERENCIADO INFILTRANTE. | | 2018 |
| | /7/2018 ADENOCARCINOMA MODERADAMENTE DIFERENCIADO INFILTRANTE Y ULCERADO. | | 2021 |
| | /7/2018 ADENOCARCINOMA MODERADAMENTE DIFERENCIADO INFILTRANTE Y ULCERADO. | | 2018 |
| | /8/2018 ADENOCARCINOMA MODERADAMENTE DIFERENCIADO INFILTRANTE Y ULCERADO. | | 2021 |
| 6 | FEMENINO 10/8/2018 ADENOCARCINOMA MODERADAMENTE DIFERENCIADO INFILTRANTE. | | 2018 |
| 7 | MASCULINO 21/12/2018 ESÓFAGO CON INVASION DE ADENOCARCINOMA. | | 2018 |
| 8 | FEMENINO 27/4/2018 ADENOCARCINOMA MODERADAMENTE DIFERENCIADO INFILTRANTE. | | 2018 |
| 9 | FEMENINO 18/7/2018 ADENOCARCINOMA MODERADAMENTE DIFERENCIADO INFILTRANTE Y ULCERADO. | | 2022 |
| 10 | FEMENINO 15/8/2018 ADENOCARCINOMA MODERADAMENTE DIFERENCIADO INFILTRANTE Y ULCERADO. | | 2021 |
| 11 | FEMENINO 16/8/2018 ADENOCARCINOMA MODERADAMENTE DIFERENCIADO INFILTRANTE. | | 2018 |
| 12 | FEMENINO 21/12/2018 ESÓFAGO CON INVASION DE ADENOCARCINOMA. | | 2018 |
| 13 | MASCULINO 21/12/2018 ESÓFAGO CON INVASION DE ADENOCARCINOMA. | | 2020 |
| 14 | FEMENINO 27/4/2018 ADENOCARCINOMA MODERADAMENTE DIFERENCIADO INFILTRANTE. | | 2018 |
| 15 | FEMENINO 18/7/2018 ADENOCARCINOMA MODERADAMENTE DIFERENCIADO INFILTRANTE Y ULCERADO. | | 2020 |
| 16 | FEMENINO 18/7/2018 ADENOCARCINOMA MODERADAMENTE DIFERENCIADO INFILTRANTE Y ULCERADO. | | 2018 |
| 17 | FEMENINO 18/7/2018 ADENOCARCINOMA MODERADAMENTE DIFERENCIADO INFILTRANTE Y ULCERADO. | | 2019 |
| 18 | FEMENINO 15/8/2018 ADENOCARCINOMA MODERADAMENTE DIFERENCIADO INFILTRANTE Y ULCERADO. | | 2018 |



Dana Smith R, Ballejo C

Quick cancer report starting from a raw database is licenced under CC BY-NC 4.0



rpc@ine.gov.ar

Introduction

This is a useful tool for the daily practice of cancer registries and has three utilities that can greatly reduce work time and strengthen the confidence of institutions. And it's free.

- It can be used to convert a set of Microsoft Word or PDF files into an ordered database, so you don't have to read each file.
- It can also be used to quickly identify cancer cases in a database, and quickly register them in the CanReg system.
- Even more, it can be used to prepare a quick report for the institutions that provide us their information.

Methods

1°. Scrapping

Input: Files (histopathological reports) .docx .pdf

Process: Read the files, scrap the files and obtain a data table

https://github.com/cballejo/basic_cancer_report/tree/main/1-report_scrapping

file: Convert docx files to csv templates.R (RStudio®) - output: raw data table (csv)

2°. Data processing

Input: raw data table (csv)

Process: Read the .csv file, detect diagnostic keywords, obtain a cancer data table

https://github.com/cballejo/basic_cancer_report/tree/main/2-data_processing

file: Database leaker.R (RStudio®) - auxiliary file: keyword detection function - output: processed data table (csv)

3° Report

Input: processed data table (csv)

Report structure:• Written introduction• Table of incident cases per year• Table of cancer cases per year and according to location site• Interannual frequency comparison graph found in the institution (top 10 primary sites)• Summary measures of cases• Frequencies of cancer cases by gender and location•Final comments

https://github.com/cballejo/basic_cancer_report/tree/main/3-final_report

file: Report.Rmd (rmarkdown R) - auxiliary files: templates and header banner - output: final word file (docx)

Language R version 4.2.0 – Rstudio 2022.07.1+554

Packages: tidyverse – officer – fs – glue – lubridate – flextable – janitor – gtsummary - cowplot

Results

A comparison was made between three reports prepared with this methodology and three reports of the same cases but once coded with CIEO-3 in CanReg5.

No differences were observed in the number of cancer cases detected and there were only 3% of cases not included in the corresponding topographical category, but in the category "other tumor locations".

Therefore, this is an effective methodology to detect cancer cases from a set of Microsoft Word or PDF files and provide institutions with a quick and useful statistical report.

FAQ + proposals + opinions

https://github.com/cballejo/basic_cancer_report/tree/main/faq