# University of Westminster

## Individual Coursework – 5DATA004C Data Science Project Lifecycle

Name: Mohammed Alfar
Student ID: w2053037/20221802
Dataset: Cancer Incidence Data from NCCP Report 2020 (Sri Lanka)

Link to Video Presentation: https://drive.google.com/file/d/1voIIT8iaFkUBV5hu2uG7dqVgwWYnz-9x/view?usp=sharing

Link to Streamlit app: https://dspl-cancer-2022--dashboard-project-mw87kigizvxncznqhfcfmj.streamlit.app/

Link to GitHub repository: https://github.com/AlfarRafeek/DSPL-Cancer-2022--Dashboard-Project/tree/main

## Table of Contents

# Data Selection

This project is using data taken from the National Cancer Control Programme (NCCP) - Annual Report: Cancer Incidence Data, Sri Lanka 2020 which is publicly accessible in PDF form from the official Ministry of Health Sri Lanka website. https://www.nccp.health.gov.lk/storage/post/pdfs/Cancer%20Control%20Programe%202023%2006%2021.qxp_Layout%201.pdf

This dataset meets the criteria of coursework requirements based on the following attributes:

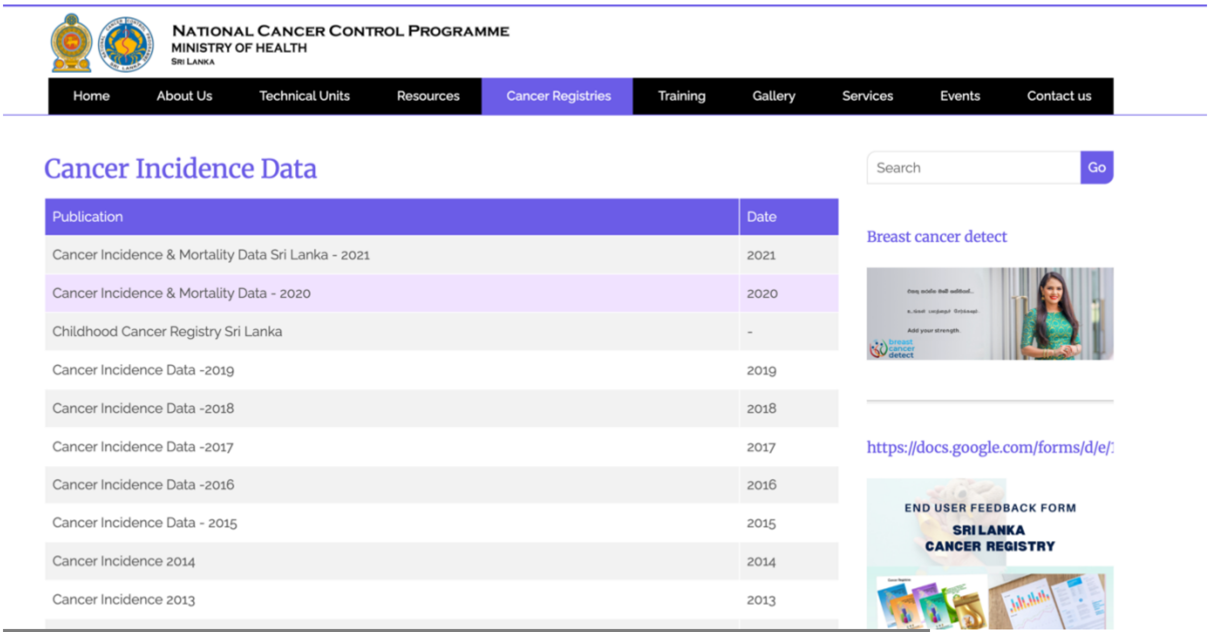
- Publicly Findable: The report is published and made freely accessible for public use by the Sri Lankan Ministry of Health.

- Single Year Level of Depth: Although it only reports data for the year 2020, it does not lose the level of depth about categorization of 198 different types of cancer (by gender) which allows for thorough analysis of specific domain.

- Represents national health trends: The data also provides invaluable knowledge into national disease trends and public health priorities which is a significant cornerstone when advising possible policy changes and collaborative awareness strategies.

- Manually Cleaned in Nature: Although the report does not represent raw dataset, it has been carefully taken from the PDF tables (tables 15 to 31), and cleaned for structure, and transformed into an Excel format usable for analysis.

This dataset is a valuable view into cancer trends and gender disparities regarding cancer in Sri Lanka, it also fits well with the specific goals in health informatics and visualization, for the purposes of this project.

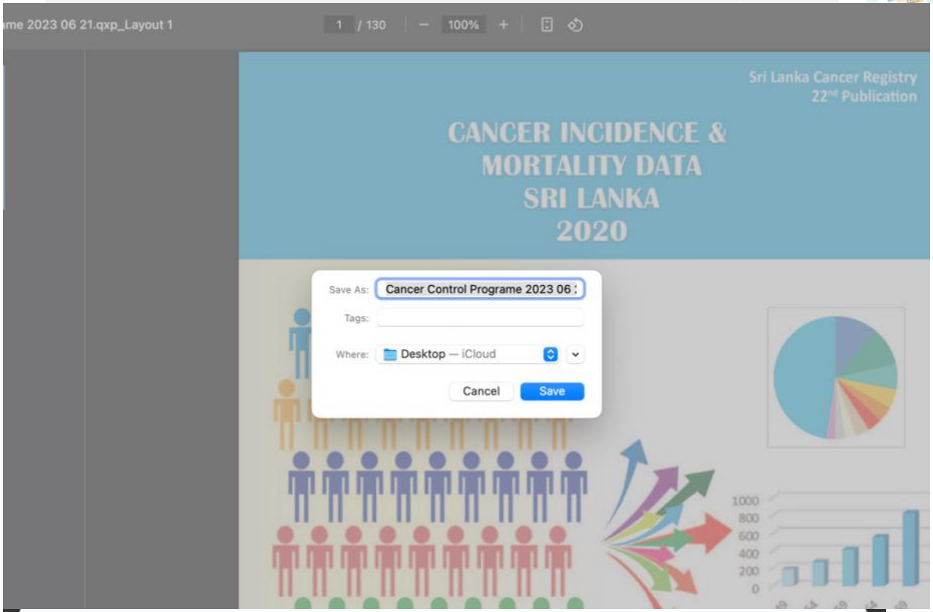


## Cancer Incidence Data

| Publication | Date |
|---|---|
| Cancer Incidence & Mortality Data Sri Lanka - 2021 | 2021 |
| Cancer Incidence & Mortality Data - 2020 | 2020 |
| Childhood Cancer Registry Sri Lanka | - |
| Cancer Incidence Data -2019 | 2019 |
| Cancer Incidence Data -2018 | 2018 |
| Cancer Incidence Data -2017 | 2017 |
| Cancer Incidence Data -2016 | 2016 |
| Cancer Incidence Data - 2015 | 2015 |
| Cancer Incidence 2014 | 2014 |
| Cancer Incidence 2013 | 2013 |

Breast cancer detect

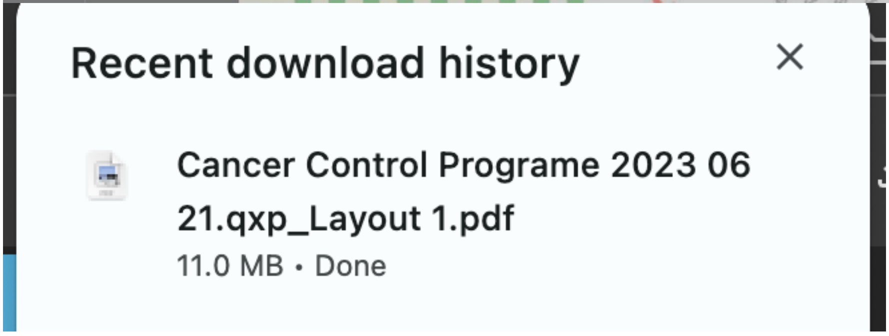
https://docs.google.com/forms/d/e/1



Save As: Cancer Control Programe 2023 06 2

Tags:

Where: Desktop — iCloud



## Recent download history

Cancer Control Programe 2023 06 21.qxp_Layout 1.pdf

11.0 MB • Done

# Aim and Objective

## Aim

The primary aim of the project was to make an interactive data visualization dashboard using Streamlit, based on cancer incidence in Sri Lanka for the year 2020. The evidence dataset was extracted from the National Cancer Control Programme (NCCP) Annual Report. The data was extracted manually from the PDF report, and it was cleaned into an Excel dataset. The cleaned dataset consisted of 198 records of cancer types broken down by gender (Male and Female).

## Objective

- To help users in understanding how cancer types differ in frequency, emphasizing, on the one hand, the most frequent cancer types and, on the other hand, the least frequent in Sri Lanka for 2020.
- To enable meaningful comparison of the incidence of cancers of males and females and understand gender differences in cancer.
- To create a data experience that enables users to search and filter the data so they can quickly locate what they're interested in concerning cancers.
- To inform insights clearly using various visual tools including bar charts, donut charts, scatter plots, and line charts.
- To not only make the dashboard functional but accessible to a broader audience for both academic purposes as well as public understanding.

# Development methodology

The data for this project was obtained from the 2020 Annual Report from the National Cancer Control Programme (NCCP) of Sri Lanka. This dataset was available only as a static PDF and was not available either publicly as a dataset through an API or as a service through the web. For these reasons, the data was extracted manually, only Tables 15-31 were reviewed, and incorporated into a new structured Excel file for accuracy and consistency. Although this method does not include automated scraping or API integration, given the official format, this was the safest and simplest approach.The final dataset includes 199 entries, with fields for cancer type, male cases, and female cases.

The dashboard was finally completed and developed using Python with libraries such as Pandas - to handle all data and Plotly - to provide interactive visualizations and Streamlit - to render the web interface). Visualizations were developed in separate units to allow for bar chart, line chart, donut chart, scatter plot potentials, in order to derive meaning. GitHub was used to version control and facilitate project movements throughout the development. The final solution was able to be publicly deployed using Streamlit Cloud. The solution was tested through functional (The programming worked correctly, to the expectations) and user centered (what the user experience was) test cases, all documented in a formal log of tests.

# Requirements

## Functional Requirements

- Users will be able to filter data by gender (Male, Female, All).
- The search function will enable users to breakdown the data types of cancer.
- The project will use multiple visual types. Including: bar, donut, line, scatter plots.
- Users will be able to see the total case counts based on their filters.
- The dashboard will be deployed and usable.

## Non-Functional Requirements

- The dashboard will take between 3-5 seconds to load.
- The interface will be professional and clean and easy to use.
- The app must be responsive on devices (desktop and mobile).
- The GitHub repo will include all the code and datasets with version control.
- The dashboard will be live and accessible until 1 August 2025 for evaluation purposes.

# Testing

## Test Log

| TC | Date | Executed by | Actual Result | Pass/Fail | Notes |
|---|---|---|---|---|---|
| TC1 | 2025-04-28 | Mohammed Alfar | Dashboard loaded all charts and elements | Pass | Working link deployed |
| TC2 | 2025-04-28 | Mohammed Alfar | Total count matched the sum in the excel file | Pass | Validated totals with calculator |
| TC3 | 2025-04-29 | Mohammed Alfar | Bar chart shows correct top 10 cancer types | Pass | Order verified with dataset |
| TC4 | 2025-04-29 | Mohammed Alfar | Donut chart shows 100% of male/female distribution | Pass | Labels correspond with proportions |
| TC5 | 2025-04-30 | Mohammed Alfar | Rare cancers listed with values under 10 in the upper limit | Pass | Filter logic verified manually clicked figures |
| TC6 | 2025-04-30 | Mohammed Alfar | Gender filter is reflecting on all charts as expected | Pass | Male/female/all filters are responding visually |
| TC7 | 2025-04-30 | Mohammed Alfar | Searching using the keyword "Breast" returns the correct chart | Pass | Exact results were visible in multilple charts |
| TC8 | 2025-05-01 | Mohammed Alfar | Streamlit sidebar navigation is nice and responsive | Pass | All controls appeared to work and quick to respond |
| TC9 | 2025-05-01 | Mohammed Alfar | Scatter plot average cancer cases between each type of cancer displayed the correct values | Pass | Spot checked values for 5 types. |
| TC10 | 2025-05-01 | Mohammed Alfar | All visualisations resized correctly on both mobile and desktop viewports | Pass | Responsiveness and layout confirmed in Chrome. |

## Test cases

| # | TC1 | Test case title: Open dashboard app |
|---|---|---|
| Description | To confirm that the Streamlit dashboard opens and correctly loads in a browser | |
| Steps / Input | 1. Open Streamlit app URL<br>2. 2. Confirm the dashboard is loading with no errors. | |
| Dependencies | Internet<br>Dashboard must be deployed and online | |
| Expected Result | The dashboard has loaded fully in the browser, with all visual components present and working. | |

| # | TC2 | Test case title: Gender filter |
|---|---|---|
| Description | To prove that selecting different gender filters (Male, Female or All), updates the visualizations accordingly. | |
| Steps / Input | 1. Open the dashboard.<br>2. On the side-bar.<br>3. Find the Gender Filter dropdown or radio buttons.<br>4. Select "Male" and confirm that the charts have updated.<br>5. Select "Female" and confirm that the charts have updated.<br>6. Select "All" ensuring that all the data returns. | |
| Dependencies | • Data must be loaded and tied to logic of gender filter.<br>• Sidebar must be active. | |
| Expected Result | Visualizations are responsive and reflect the gender category selected. | |

| # | TC3 | Test case title: Search Cancer Type |
|---|---|---|
| Description | Confirm that the search option filters cancer types based on user input. | |
| Steps / Input | 1. Load the dashboard and ensure it loads completely.<br>2. Locate the Search Cancer Type textbox in the sidebar.<br>3. Type in a search term, "Lung".<br>4. Watch for changes in all charts and indicators. | |
| Dependencies | • The sidebar and input field must be loaded and functional.<br>• The search function must be coded correctly. | |
| Expected Result | Chart automatically filter to represent only cancer types matching the search term. | |

| # | TC4 | Test case title: Top 10 Cancer Types Chart |
|---|---|---|
| Description | Verify that the Top 10 Cancer Types bar chart is loaded correctly and shows data you expect. | |
| Steps / Input | 1. Load the dashboard completely.<br>2. Ensure Gender filter is set to 'All'.<br>3. Scroll to find (or locate) the Top 10 Cancer Types horizontal bar chart.<br>4. Review the bar chart's titles, labels, and values. | |
| Dependencies | • Filter options must be functional.<br>• Chart library (Plotly) must be working correctly. | |
| Expected Result | A horizontal bar chart shows the Top 10 cancer types ranked by number of cases. | |

| # | TC5 | Test case title: RARE CANCER TYPES (less than 10) |
|---|---|---|
| Description | Verify the rare cancer types of charts (fewer than 10 cases) appears and filters appropriately. | |
| Steps / Input | 1. Open the dashboard.<br>2. Scroll to the Rare Cancer Types or Types with <10 Cases section.<br>3. verify the chart garners logic as expected.<br>4. Verify there are no types with ≥10 cases shown. | |
| Dependencies | • Must have data filter for count less than 10.<br>• The data must be fully loaded first. | |
| Expected Result | A chart shows only cancer types with under 10 recorded cases. | |

# Conclusion

The primary aims of this project to develop an interactive Streamlit dashboard for Sri Lanka's cancer incidence data from 2020 were achieved. I employed several technologies to support my work on the dashboard, and they were quite successful in allowing me to develop effective visualizations to transform public health data into useful knowledge. The dashboard hosts a simple and interactive web interface for displaying the incidence rate of cancer among different groups, by site and by rarity of the cancer. The use of Python, Streamlit and GitHub allowed us to create a contemporary and extensible implementation and deployment stack.

The manual extraction and cleaning of the NCCP data from their PDF report was more challenging than expected. This step required us to consider the need to focus on both accuracy and structure. Future work could eliminate time and exponentially expand scalability by searching for official datasets that are in organized formats or by automatically gathering data from official sources.

overall, this project was a valuable learning opportunity in data wrangling, data visualization and deployment. this project shows what can be accomplished through data science to support public health awareness, along with decision making. the dashboard was made with considerations of both the academic reviews as well as intended special outputs for general users looking to get insights into cancer trends in Sri Lanka.