1. Overall description

1.1 Datasets preparation

A total of 14 datasets were prepared for analysis after the data wrangling process, categorized into two groups: cancer data and risk factors, which are stored in the 'data/clean' folder on GitHub. The cancer data encompasses 'incidence' and 'mortality' records. The risk factors are further classified into two categories: environmental factors, including 'earthquakes', 'air quality', 'groundwater quality', and 'temperature', and human factors, including 'income,' 'highest qualification level', 'working hours' and "number of children'. The cancer data spans from 2011 to 2020, environmental factors are documented for the same period, while human factors are specifically available for the years 2013 and 2018.

1.2 Common dataset information

The first two columns consist of 'year' and 'DHB' in all the datasets, documenting specific year-based information along with the corresponding DHB (District Health Board) regions. In the context of correlation analysis, the cancer data and risk factors are linked using 'DHB' and 'year' as common identifiers. Each year within each DHB region is treated as an individual sample. For instance, with a total of 20 regions, there are 20 samples for each year. Therefore, when considering cancer and environmental data, there are 200 samples spanning from 2011 to 2020 for each dataset. For human factors, there are 40 samples available, covering the years 2013 and 2018 for each dataset. Apart from these common identifiers, all other column names in the risk factor data are structured as 'dataset name-variable'. For instance, in the 'earthquake' dataset, the column for recording the highest earthquake magnitude is named 'earthquake-magnitude_max.'

2 Datasets

2.1 Incidence

2.1.1 Format

Original dataset: CSV file

Cleaned dataset: CSV file

2.1.2 Data source

Cancer web tool: "This web tool presents cancer registrations data from the New Zealand Cancer Registry and cancer deaths data from the New Zealand Mortality Collection. Both are held by Te Whatu Ora – Health New Zealand. Cancer registration data was extracted on 11 January 2023 and cancer death data was extracted on 26 October 2022" (Cancer web tool).

2.1.3 Accessed website

Official website: https://tewhatuora.shinyapps.io/cancer-web-tool/

Data accessed website**:** https://tewhatuora.shinyapps.io/cancer-web-tool/

2.1.4 Description

This dataset includes the information of the incidence number and rate, and relevant gender for 14 different types of cancer in each DHB region from 2011 to 2020. The breast, ovarian, thyroid, and uterine cancers are exclusive to females, while bladder, kidney, prostate and testicular cancers are exclusive to males. The other cancer types encompass data for both males and females.

2.1.5 Structure

6301 rows and 6 columns

2.1.6 Dataset variables

Variables	Description
DHB	The region of District Health Board
year	The year for which cancer incidence data is recorded, from 2011 to 2020
sex	The gender of the cancer incidence, including male, female, and all sex. All sex represents the combined data for both genders. If a specific cancer contains one gender, the "all sex" category would include data relevant to that particular gender.
cancer	The types of cancer, including 14 different types
incidence_num	The number of the cancer incidence
incidence_rate	The rate of the cancer incidence (‰)

2.2 Incidence_sexfiltered

2.2.1 Format

Original dataset: CSV file

Cleaned dataset: CSV file

2.2.2 Data source

Cancer web tool

2.2.3 Accessed website

Official website: https://tewhatuora.shinyapps.io/cancer-web-tool/

Data accessed website**:** https://tewhatuora.shinyapps.io/cancer-web-tool/

2.2.4 Description

This dataset is another version of ‘incidence’ but with cancers grouped by gender. More specifically, for cancers that exclusively affect either males or females, their data are classified either male or female group, while cancers that have the potential to affect both males and females are grouped into ‘all sex’.

2.2.5 Structure

2941 rows and 7 columns

2.2.6 Dataset variables

Variables	Description
DHB	The region of District Health Board
year	The year for which cancer incidence data is recorded, from 2011 to 2020
sex	The gender of the cancer incidence data, including male, female, and all sex.
cancer	The types of cancer, including 14 different types
incidence_num	The number of the cancer incidence
incidence_rate	The rate of the cancer incidence (‰)
group	The gender grouping of cancer types, including male, female, and all sex

2.3 Mortality

2.3.1 Format

Original dataset: CSV file

Cleaned dataset: CSV file

2.3.2 Data source

Cancer web tool

2.11.3 Accessed website

Official website: https://tewhatuora.shinyapps.io/cancer-web-tool/

Data accessed website**:** https://tewhatuora.shinyapps.io/cancer-web-tool/

2.3.4 Description

This dataset represents the mortality number and rate, and gender information for 14 different types of cancers in each DHB from 2011 to 2020.

2.3.5 Structure

6301 rows and 6 columns

2.3.6 Dataset variables

Variables	Description
DHB	The region of District Health Board
year	The year for which cancer mortality data is recorded, from 2011 to 2020
sex	The gender of the mortality data, including male, female, and all sex.
cancer	The types of cancer, including 14 different types
mortality_num	The number of mortalities
mortality_rate	The rate of mortality

2.4 Mortality_sexfiltered

2.4.1 Format

Original dataset: CSV file

Cleaned dataset: CSV file

2.4.2 Data source

Cancer web tool

2.4.3 Accessed website

Official website: https://tewhatuora.shinyapps.io/cancer-web-tool/

Data accessed website**:** https://tewhatuora.shinyapps.io/cancer-web-tool/

2.4.4 Description

This dataset is another version of 'mortality' dataset, but with cancers grouped by gender. The classification rules are the same as the 'Incidence_sexfiltered' dataset.

2.4.5 Structure

2941 rows and 7 columns

2.4.6 Dataset variables

Variables	Description
DHB	The region of District Health Board
year	The year for which cancer mortality data is recorded, from 2011 to 2020
sex	The gender of the mortality data, including male, female, and all sex.

cancer	The types of cancer, including 14 different types
mortality_num	The number of mortalities
mortality_rate	The rate of mortality
group	The gender grouping of cancer types, including male, female, and all sex

2.5 Earthquake

2.5.1 Format

Original dataset: CSV file

Cleaned dataset: CSV file

2.5.2 Data source

GeoNet: a partnership involving EQC Toka Tū Ake (Natural Hazards Commission) and GNS Science (Institute of Geological and Nuclear Sciences Limited) (GeoNet).

2.5.3 Accessed website

Official website: https://www.geonet.org.nz/

Data accessed website: https://quakesearch.geonet.org.nz/

2.5.4 Description

This dataset covers earthquakes in New Zealand from 2011 to 2020, providing information on their magnitude, depth, and counts. In the original dataset, geographic location information is recorded using longitude and latitude in decimal degrees, with the coordinates following the WGS84 datum. We've geospatially matched these coordinates to specific corresponding DHB region. For each year, we've calculated the maximum magnitude, average magnitude, maximum depth, average depth, and the frequency of earthquakes in each DHB region.

2.5.5 Structure

201 rows and 7 columns

2.5.6 Dataset variables

Variable Name	Description
Year	The year earthquakes occurred, from 2011 to 2020.
DHB	The region of District Health Board
Earthquake-magnitude_max	The highest earthquake magnitude in a specific year.
Earthquake-magnitude_mean	The average earthquake magnitude in a specific year.
Earthquake-depth_max	The highest earthquake depth in a specific year.
Earthquake-depth_mean	The average earthquake depth in a specific year.
Earthquake-counts	The frequency of earthquakes in a specific year.

2.6 Air

2.6.1 Format

Original dataset: XLSX file

Cleaned dataset: CSV file

2.6.2 Data source

LAWA: “LAWA (Land, Air, Water Aotearoa) has been established by like-minded organisations with a view to helping local communities find the balance between using natural resources and maintaining their quality and availability. LAWA is now a partnership between the Te Uru Kahika - Regional and Unitary Councils Aotearoa, Cawthron Institute, the Ministry for the Environment, the Department of Conservation, Stats NZ and has been supported by the Tindall Foundation and Massey University" (LAWA).

2.6.3 Accessed website

Official website: https://www.lawa.org.nz/

Data accessed website: https://www.lawa.org.nz/download-data/

2.6.4 Description

The original data file published on June 21, 2023, recording concentrations of PM10 (particles with a diameter less than 10 µm) and PM2.5 (particles with a diameter less than 2.5 µm) from air quality monitoring sites across New Zealand. We converted the geographical information to distinct DHB regions according to latitude and longitude. Following this transformation, we computed both the highest and average PM10 and PM2.5 concentrations for each region on an annual basis. According to LAWA, PM10 and PM2.5 are types of airborne particles that can have adverse health effects. PM10 particles can enter our respiratory, while PM2.5 particles can penetrate deep into our lungs.

2.6.5 Structure

89 rows and 6 columns

2.6.6 Dataset variables

Variables	Description
year	The year for which air quality data is recorded, from 2011 to 2020.
DHB	The region of District Health Board
Air-concentration_max_PM10	The highest concentration of PM10
Air-concentration_max_PM2.5	The highest concentration of PM2.5
Air-concentration_mean_PM10	The average concentration of PM10
Air-concentration_mean_PM2.5	The average concentration of PM2.5

2.7 Water

2.7.1 Format

Original dataset: XLSX file

Cleaned dataset: CSV file

2.7.2 Data source

LAWA

2.7.3 Accessed website

Official website: https://www.lawa.org.nz/

Data accessed website: https://www.lawa.org.nz/download-data/

2.7.4 Description

The original data file published on November 24, 2022, recording the ground water quality monitoring by New Zealand's regional councils and unitary authorities. Groundwater serves as a vital source of fresh water for various purposes in New Zealand, including drinking water, irrigation, industrial use, and the sustenance of numerous streams and lakes (LAWA). There are five indicators for the quality of ground water in this dataset, each of which includes the maximum and average values for each DHB region annually. The DHB region information was derived from the latitude and longitude data in the original dataset.

2.7.5 Structure

201 rows and 12 columns

2.7.6 Dataset variables

Variables	Description
year	The year for which ground water quality data is recorded, from 2011 to 2020.
DHB	The region of District Health Board
Water-censoredValue_max_Chloride	The highest value of the chloride (g/mÂ³)
Water-censoredValue_max_Dissolved Reactive Phosphorus	The highest value of the dissolved reactive phosphorus (g/mÂ³)
Water-censoredValue_max_E. Coli	The highest value of the E. Coli (CFU/100ml)
Water-censoredValue_max_Electrical Conductivity	The highest value of the electrical conductivity (ÂµS/cm)
Water-censoredValue_max_Nitrate Nitrogen	The highest value of the nitrate nitrogen (g/mÂ³)
Water-censoredValue_mean_Chloride	The average value of the chloride (g/mÂ³)
Water-censoredValue_mean_Dissolved Reactive Phosphorus	The average value of the dissolved reactive phosphorus (g/mÂ³)
Water-censoredValue_mean_E. Coli	The average value of the E. Coli (CFU/100ml)
Water-censoredValue_mean_Electrical Conductivity	The average value of the electrical conductivity (ÂµS/cm)
Water-censoredValue_mean_Nitrate Nitrogen	The average value of the nitrate nitrogen (g/mÂ³)

2.8 Temperature

2.8.1 Format

Original dataset: CSV file

Cleaned dataset: CSV file

2.8.2 Data source

Statistic NZ: the official data agency of New Zealand, gathering data from individuals and organizations via censuses and surveys (Stats NZ).

2.8.3 Accessed website

Official website: https://www.stats.govt.nz/

Data accessed website: https://www.stats.govt.nz/indicators/temperature/

2.8.4 Description

This dataset contains the annual and seasonal temperature trends from 2011 to 2020, organized by DHB regions, including the highest and average temperature by Celsius degree. The DHB region information was derived by converting the latitude and longitude data from the original dataset.

2.8.5 Structure

181 rows and 17 columns

2.8.6 Dataset variables

Variables	Description
year	The year for which temperature data is recorded, from 2011 to 2020.
DHB	The region of District Health Board
Temperature-Average_Annual	The annual average temperature
Temperature-Average_Autumn	The average temperature in Autumn
Temperature-Average_Spring	The average temperature in Spring
Temperature-Average_Summer	The average temperature in Summer
Temperature-Average_Winter	The average temperature in Winter
Temperature-Maximum_Annual	The annual highest temperature
Temperature-Maximum_Autumn	The highest temperature in Autumn
Temperature-Maximum_Spring	The highest temperature in Spring
Temperature-Maximum_Summer	The highest temperature in Summer
Temperature-Maximum_Winter	The highest temperature in Winter
Temperature-Minimum_Annual	The annual lowest temperature
Temperature-Minimum_Autumn	The lowest temperature in Autumn
Temperature-Minimum_Spring	The lowest temperature in Spring
Temperature-Minimum_Summer	The lowest temperature in Summer
Temperature-Minimum_Winter	The lowest temperature in Winter

2.9 Work_hours

2.9.1 Format

Original dataset: CSV file

Cleaned dataset: CSV file

2.9.2 Data source

Statistic NZ

2.9.3 Accessed website

Official website: https://www.stats.govt.nz/

Data accessed website: https://www.stats.govt.nz/information-releases/statistical-area-1-dataset-for-2018-census-updated-march-2020

2.9.4 Description

This dataset provides information on the working hours of the population proportion within each DHB region. The data is based on the 2013 and 2018 Censuses, offering insights into the regional workforce. In the original dataset, geographic information was recorded using area coded. We have matched this information to the corresponding DHB regions.

2.5.5 Structure

43 rows and 15 columns

2.5.6 Dataset variables

Variable Name	Description
DHB	The region of District Health Board
year	The specific year for the work hour records, including 2013 and 2018.
WorkHours-1-9 hours worked	The proportion (%) of the population that worked between 1 and 9 hours.
WorkHours-10-19 hours worked	…
WorkHours-20-29 hours worked	…
WorkHours-30-39 hours worked	…
WorkHours-40-49 hours worked	…
WorkHours-50-59 hours worked	…
WorkHours-60 hours or more worked	The proportion (%) of the population that worked 60 hours or more.
WorkHours-≥10 hours	The proportion (%) of the population that worked equal or greater than 10 hours.
WorkHours-≥20 hours	…
WorkHours-≥30 hours	…
WorkHours-≥40 hours	…
WorkHours-≥50 hours	…
WorkHours-≥60 hours	The proportion (%) of the population that worked equal or greater than 60 hours.

2.10 Highest_qualification

2.10.1 Format

Original dataset: CSV file

Cleaned dataset: CSV file

2.10.2 Data source

Statistic NZ

2.10.3 Accessed website

Official website: https://www.stats.govt.nz/

Data accessed website**:** https://www.stats.govt.nz/information-releases/statistical-area-1-dataset-for-2018-census-updated-march-2020

2.10.4 Description

This dataset provides information on the proportion of the population with the highest qualification for each DHB region in 2013 and 2018. The DHB region data was derived from the conversion of area codes from the original dataset, which is based on the 2013 and 2018 Census. We classified the qualification level in accordance with New Zealand's standards as following (careers.govt.nz):

Level 1 certificates

Level 2 certificates

Level 3 certificates

Level 4 certificates

Level 5 certificates and diplomas

Level 6 certificates and diplomas
Level 7 graduate certificates, graduate diplomas and Bachelor's degrees
Level 8 postgraduate certificates, postgraduate diplomas and Bachelor's Honours degrees
Level 9 Master's degrees
Level 10 doctoral degrees.

2.10.5 Structure

43 rows and 23 columns

2.10.6 Dataset variables

Variables	Description
DHB	The region of District Health Board
year	The specific year for the work hour records, including 2013 and 2018.
Education-No qualification	The proportion (%) of population without qualification
Education-level 1	The proportion (%) of population with level 1 qualification
Education-level 2	…
Education-level 3	…
Education-level 4	…
Education-level 5	…
Education-level 6	…
Education-level 7	…
Education-level 8	…
Education-level 9	…
Education-level 10	The proportion (%) of population with level 10 qualification
Education-≥level 1	The proportion (%) of population with qualification equal and greater than level 1
Education-≥level 2	…
Education-≥level 3	…
Education-≥level 4	…
Education-≥level 5	…
Education-≥level 6	…
Education-≥level 7	…
Education-≥level 8	…
Education-≥level 9	…
Education-≥level 10	The proportion of population with qualification equal and greater than level 10

2.11 Income

2.11.1 Format

Original dataset: CSV file

Cleaned dataset: CSV file

2.11.2 Data source

Statistic NZ

2.11.3 Accessed website

Official website: https://www.stats.govt.nz/

Data accessed website**:** https://www.stats.govt.nz/information-releases/statistical-area-1-dataset-for-2018-census-updated-march-2020

2.11.4 Description

This dataset contains the population proportions within each income level for each DHB region in both 2013 and 2018. The area codes in the original dataset are converted to the corresponding DHB regions.

2.11.5 Structure

43 rows and 15 columns

2.11.6 Dataset variables

Variables	Description
DHB	The region of District Health Board
year	The specific year for the income records, including 2013 and 2018.
Income-$5,000 or less	The population proportion (%) of income equal or less than $5000
Income-$5,001-$10,000	The population proportion (%) of income ranging from $5001 to $10,000
Income-$10,001-$20,000	…
Income-$20,001-$30,000	…
Income-$30,001-$50,000	…
Income-$50,001-$70,000	…
Income-$70,001 or more	The population proportion (%) of income equal or more than $70,001
Income-> $5,000	The population proportion (%) of income more than $5,000
Income-> $10,000	…
Income-> $20,000	…
Income-> $30,000	…
Income-> $50,000	…
Income-> $70,000	The population proportion (%) of income more than $70,000

2.12 Number of children

2.12.1 Format

Original dataset: CSV file

Cleaned dataset: CSV file

2.12.2 Data source

Statistic NZ

2.12.3 Accessed website

Official website: https://www.stats.govt.nz/

Data accessed website**:** https://www.stats.govt.nz/information-releases/statistical-area-1-dataset-for-2018-census-updated-march-2020

2.10.4 Description

This dataset presents the number of children born to each woman aged 15 and above, based on the 2013 and 2018 Censuses. It is intended for use in conducting correlation analysis related to cancers specific to females. The area codes in original dataset have been converted into corresponding DHB regions.

2.12.5 Structure

43 rows and 15 columns

2.12.6 Dataset variables

Variables	Description
DHB	The region of District Health Board
year	The specific year for the number of children records, including 2013 and 2018.
Children-No children	The proportion (%) of population with no children.
Children-One child	…
Children-Two children	…
Children-Three children	…
Children-Four children	…
Children-Five children	…
Children-Six or more children	The proportion (%) of population with six or more children
Children-> 0 children	The proportion (%) of population with at least one child.
Children-> 1 children	The proportion (%) of population with more than one child (not included).
Children-> 2 children	…
Children-> 3 children	…
Children-> 4 children	…
Children-> 5 children	The proportion (%) of population with more than five children (not included).