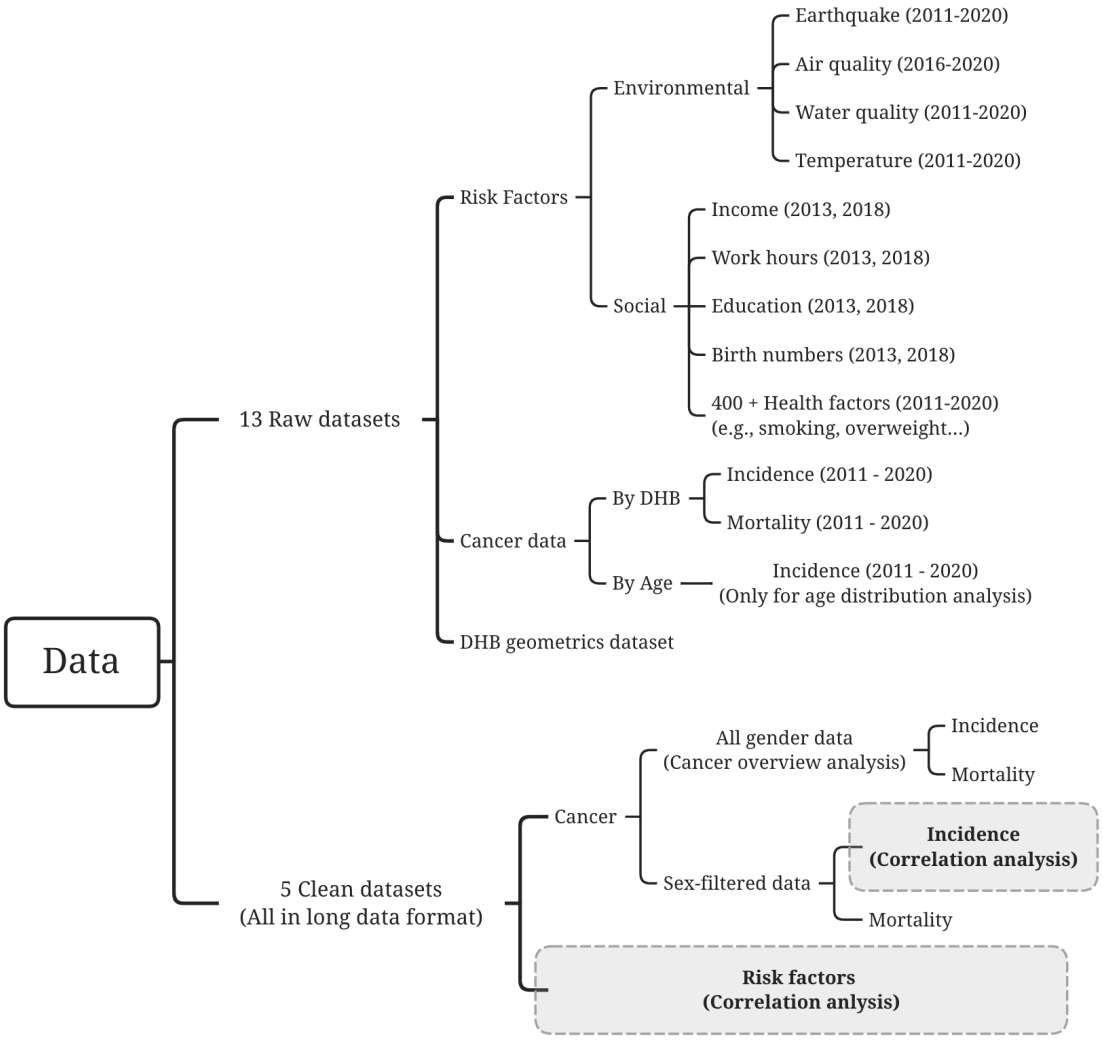
1. Data Overview

Raw data: A total of 13 raw datasets are used this project: 3 cancer datasets, 9 risk factors datasets, and 1 District Health Board (DHB) geometrics dataset. All raw datasets are saved in "data/raw".

Clean data: There are 5 clean datasets after data wrangling, which are saved in "data/clean". All risk factors are combined in a clean long data frame. Cancer incidence and mortality are separated into two different datasets.

An overview of all datasets is shown as followings:



Clean sex-filtered cancer dataset: primary key {DHB, year, sex, cancer}

Clean risk factors dataset: primary key {DHB, year, sex, category, rf}

As shown in the grey box above, "Clean sex-filtered cancer incidence dataset" and "clean risk factors dataset" are used for correlation analysis. For sex-filtered cancer datasets, there is only one sex category for each cancer type. For example, "All sex" for "Lung cancer", "Female" for "Breast cancer", "Male" for "Prostate cancer". For risk factors, all risk factors have single sex category "All sex", except NZHS risk factors, which has different sex categories (All sex/ Female / Male). Therefore, when

connecting cancer data to risk factors data, common identifier is {DHB, year}, except NZHS risk factors, where common identifier is {DHB, year, sex}.

2. Raw Datasets

2.1 Cancer Incidence by DHB

Cancer Incidence by DHB	
File name	cancer-registrations-by-dhb.csv
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, 2023).
Download link	https://tewhatuora.shinyapps.io/cancer-web-tool/
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.

2.2 Cancer Mortality by DHB

Cancer Mortality by DHB	
File name	cancer-deaths-by-dhb.csv
Data Source	Same as "Cancer Incidence by DHB"
Download link	https://tewhatuora.shinyapps.io/cancer-web-tool/
Description	This dataset represents the mortality number and rate, and relevant gender information for 14 different types of cancers in each DHB region from 2011 to 2020. The description of 'sex' is the same as the 'incidence' dataset.

2.3 Cancer Incidence by Age

Cancer Incidence by Age	
File name	cancer-registrations-by-age.csv
Data Source	Same as "Cancer Incidence by DHB"
Download link	https://tewhatuora.shinyapps.io/cancer-web-tool/
Description	This dataset includes the information of the overall cancer incidence number and rate for different age group. **Please note:** Cancer types and DHB information are unavailable.

2.4 DHB geometrics dataset

DHB geometrics dataset	
File name	NZ_District_Health_Board_boundaries_-_generalised.kml
Data Source	Statistic NZ: the official data agency of New Zealand, gathering data from individuals and organizations via censuses and surveys (Statistics NZ, n.d.).
Download link	https://datafinder.stats.govt.nz/layer/87883-district-health-board-2015/
Description	This dataset includes geometrics information for each DHB regions, which is used for region mapping with coordinates information in environmental risk factors datasets.

2.5 Earthquake dataset

Earthquake dataset	
File name	earthquake2007-2023.csv
Data Source	GeoNet: a partnership involving EQC Toka Tū Ake (Natural Hazards Commission), GNS Science (Institute of Geological and Nuclear Sciences Limited), and LINZ (Land Information New Zealand) (GeoNet, n.d.).
Download link	https://www.geonet.org.nz/
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 highest and average value for magnitude and depth, as well as the frequency of earthquakes in each DHB region.

2.6 Air quality dataset

Air quality dataset	
File name	airqualitydownloaddata_2016-2022.xlsx
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, n.d.).

Download link	https://www.lawa.org.nz/media/5261861/airqualitydownloaddata_2016-2022.xlsx
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 (LAWA, 2023). 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.

2.7 Water quality dataset

Water quality dataset	
File name	gwqmonitoringresults_sept2022.xlsx
Data Source	LAWA (Same as Air quality dataset)
Download link	https://www.lawa.org.nz/media/5261751/gwqmonitoringresults_sept2022.xlsx
Description	The original data file published on November 24, 2022, recording the ground water quality monitoring by New Zealand's regional councils. 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 from 2011 to 2020. The DHB region information was derived from the latitude and longitude data in the original dataset.

2.8 Temperature dataset

Temperature dataset	
File name	gwqmonitoringresults_sept2022.xlsx
Data Source	Statistic NZ
Download link	https://www.stats.govt.nz/assets/Uploads/Environment-indicators-2023/Temperature-indicator/Download-data/temperature-data-to-2022.zip
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.

2.9 Work Hours dataset

Work hours dataset	
File name	total_hours_worked_long_updated_16-7-20.csv
Data Source	Statistic NZ
Download link	https://www3.stats.govt.nz/2018census/SA1Dataset/Statistical%20Area%201%20dataset%20for%20Census%202018%20%E2%80%93%20total%20New%20Zealand%20%E2%80%93%20Long%20format_updated_16-7-20.zip?_ga=2.236684743.718705924.1697520523-2112202071.1695702209
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. In the original dataset, geographic information was recorded using area codes. We have matched this information to the corresponding DHB regions.

2.10 Education dataset

Education dataset	
File name	Highest_qualification_long_updated_16-7-20.csv
Data Source	Statistic NZ
Download link	Same as "**Work hours dataset**"
Description	This dataset provides information on the proportion of the population with the highest educational 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. We classified the qualification level in accordance with New Zealand's standards as following (careers.govt.nz, n.d.): 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

2.11 Income dataset

Income dataset	
File name	Total_personal_income_long_updated_16-7-20.csv
Data Source	Statistic NZ
Download link	Same as "**Work hours dataset**"
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.

2.12 Birth number

Birth number dataset	
File name	Number_of_children_born_long_updated_16-7-20.csv
Data Source	Statistic NZ
Download link	Same as "**Work hours dataset**"
Description	This dataset presents the number of children born to females 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.

2.13 New Zealand Health Survey

Birth number dataset	
File name	nz-health-survey-2017-20-regional-update-rgc-prevalences.csv
Data Source	Minister of Health: The government department responsible for overseeing and managing the country's healthcare and public health system.
Download link	https://minhealthnz.shinyapps.io/nz-health-survey-2017-20-regional-update/_w_27e6298c/_w_79b5c551/data/nz-health-survey-2017-20-regional-update-rgc-comparisons.csv
Description	This is a survey conducted by the New Zealand Ministry of Health. According to the Ministry of Health (2021), the surveyors were randomly selected from households in designated areas, including "one adult aged 15 years or older and one child aged 14 years or younger (if any in the household)". This dataset is categorized by DHB and covers the period from 2011 to 2019. It primarily comprises surveys on health behaviors or health status, such as smoking habits, dental health, physical activity, drinking habits, etc. This dataset involves various variables, and we list some of them that are relevant to our analysis. For the remaining variable descriptions, please refer to the data source website.

3. Clean Data

3.1 Cancer dataset (All gender)

Cancer dataset	
File name	[1] "incidence.csv" [2] "mortality.csv"
Data format	Long dataframe
Variable	
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. **Please Note:** mortality and incidence dataset have different cancer types.
incidence_num (mortality_num)	The number of the cancer registration/death
incidence_rate (mortality_rate)	Population standardized incidence/mortality rate (per 100,000 people)

3.2 Cancer dataset (Sex-filtered)

Cancer dataset	
File name	[1] "incidence_sexfiltered.csv" [2] "mortality_sexfiltered.csv"
Data format	Long dataframe
Variable	
DHB	The region of District Health Board
year	The year for which cancer incidence data is recorded, from 2011 to 2020
sex	There is only one sex category for each cancer type. For example, “AllSex” for “Lung cancer”, “Female” for “Breast cancer”, “Male” for “Prostate cancer”
cancer	The types of cancer, including 14 different types. **Please Note:** mortality and incidence dataset have different cancer types.
incidence_num (mortality_num)	The number of the cancer registration/death
incidence_rate (mortality_rate)	Population standardized incidence/mortality rate (per 100,000 people)

3.3 Combined risk factors dataset

Cancer dataset	
File name	rf.Rdata
Data format	Long dataframe
Variable	
DHB	The region of District Health Board
year	2011-2020 for Earthquake, Water, Temperature, NZHS 2013, 2018 for Income, Education, Work hours, Birth number 2016-202 for Air quality
sex	“AllSex / Male / Female” for NZHS risk factors “AllSex” for all other risk factors category
category	There are 9 categories: [1] "NZHS" [2] "Work Hours" [3] "Birth Number" [4] "Income" [5] "Education" [6] "Earthquake" [7] "Temperature" [8] "Water quality" [9] "Air quality"
rf	Detailed risk factors within each category
value	Value for risk factors, either in percentage or actual value
type	Specify types of value: percentage or actual value

3.4 Supplementary description of detailed risk factors in rf.Rdata

Supplementary table		
Category	**Variables**	**Description**
Earthquake	magnitude_max	The highest earthquake magnitude in a specific year.
	magnitude_mean	The average earthquake magnitude in a specific year.
	depth_max	The highest earthquake depth in a specific year.
	depth_mean	The average earthquake depth in a specific year.
	counts	The frequency of earthquakes in a specific year.
Air quality	PM10_concentration_max	The highest concentration of PM10
	PM10_concentration_mean	The average concentration of PM10
	PM2.5_concentration_max	The highest concentration of PM2.5
	PM2.5_concentration_mean	The average concentration of PM2.5

Temperature	Average_Annual	The annual average temperature
	Average_Autumn	The average temperature in Autumn
	Average_Spring	…
	Average_Summer	…
	Average_Winter	The average temperature in Winter
	Maximum_Annual	The annual highest temperature
	Maximum_Autumn	The highest temperature in Autumn
	Maximum_Spring	…
	Maximum_Summer	…
	Maximum_Winter	The highest temperature in Winter
	Minimum_Annual	The annual lowest temperature
	Minimum_Autumn	The lowest temperature in Autumn
	Minimum_Spring	…
	Minimum_Summer	…
	Minimum_Winter	The lowest temperature in Winter
Water quality	Chloride_max	The highest value of the chloride (g/mÂ³)
	Dissolved Reactive Phosphorus_max	…(g/mÂ³)
	E. Coli_max	…(CFU/100ml)
	Electrical Conductivity_max	…(ÂµS/cm)
	Nitrate Nitrogen_max	The highest value of the Nitrate Nitrogen (g/mÂ³)
	Chloride_mean	The average value of the Chloride (g/mÂ³)
	Dissolved Reactive Phosphorus_mean	…(g/mÂ³)
	E. Coli_mean	…(CFU/100ml)
	Electrical Conductivity_mean	…(ÂµS/cm)
	Nitrate Nitrogen_mean	The average value of the Nitrate Nitrogen (g/mÂ³)
Work Hours	1-9 hours worked	The proportion (%) of the population that worked between 1 and 9 hours.
	10-19 hours worked	…
	20-29 hours worked	…
	30-39 hours worked	…
	40-49 hours worked	…
	50-59 hours worked	…
	60 hours or more worked	The proportion (%) of the population that worked 60 hours or more.
	≥10 hours	The proportion (%) of the population that worked equal or greater than 10 hours.

	≥20 hours	…
	≥30 hours	…
	≥40 hours	…
	≥50 hours	…
	≥60 hours	The proportion (%) of the population that worked equal or greater than 60 hours.
Education	No qualification	The proportion (%) of population without qualification
	level 1	The proportion (%) of population with level 1 qualification
	level 2	…
	level 3	…
	level 4	…
	level 5	…
	level 6	…
	level 7	…
	level 8	…
	level 9	…
	level 10	The proportion (%) of population with level 10 qualification
	≥level 1	The proportion (%) of population with qualification equal and greater than level 1
	≥level 2	…
	≥level 3	…
	≥level 4	…
	≥level 5	…
	≥level 6	…
	≥level 7	…
	≥level 8	…
	≥level 9	…
	≥level 10	The proportion of population with qualification equal and greater than level 10
Income	$5,000 or less	The population proportion (%) of income equal or less than $5000
	$5,001-$10,000	The population proportion (%) of income ranging from $5001 to $10,000
	$10,001-$20,000	…
	$20,001-$30,000	…
	$30,001-$50,000	…
	$50,001-$70,000	…

	$70,001 or more	The population proportion (%) of income equal or more than $70,001
	> $5,000	The population proportion (%) of income more than $5,000
	> $10,000	…
	> $20,000	…
	> $30,000	…
	> $50,000	…
	> $70,000	The population proportion (%) of income more than $70,000
Birth number	No children	The proportion (%) of population with no children.
	One child	…
	Two children	…
	Three children	…
	Four children	…
	Five children	…
	Six or more children	The proportion (%) of population with six or more children
	> 0 children	The proportion (%) of population with at least one child.
	> 1 children	The proportion (%) of population with more than one child (not included).
	> 2 children	…
	> 3 children	…
	> 4 children	…
	> 5 children	The proportion (%) of population with more than five children (not included).
NZHS	Past-year drinkers	The population proportion (%) of drinking last year.
	Heavy episodic drinking at least monthly (total population)	The proportion (%) of total population who has heavy drinking monthly
	Heavy episodic drinking at least weekly (total population)	…
	Hazardous drinkers (total population)	The proportion (%) of total population with hazardous drinking patterns
	Heavy episodic drinking at least monthly (past-year drinkers)	The proportion (%) of past year drinkers who has heavy drinking monthly
	Heavy episodic drinking at least weekly (past-year drinkers)	…
	Hazardous drinkers (past-year drinkers)	The proportion (%) of past year drinkers with hazardous drinking patterns

	Private health insurance	The population proportion who has private health insurance
	Little or no physical activity	… has little or no physical activity
	Current smokers	… people who are currently smoking
	Daily smokers	… who are smoking daily
	Heavy smokers	… who are heavily smoking
	Obese	… who are obese
	Only visit dental health care worker for problems	… who visit the dentist only when they have a dental issue
	Dental health care worker visit	… who visit the dentist routine dental checkup
	Diabetes	… who have diabetes
	All teeth removed due to decay	… who removed all the teeth because of decay
	Mean diastolic blood pressure (mmHg)	The mean value of diastolic blood pressure (mmHg)
	Mean height (cm)	…height (cm)
	…	…
	* There are over 100 risk factors in NZHS, we only list significant ones identified in our analysis.	

References

Careers.govt.nz. (n.d.). *Qualifications and their levels* https://www.careers.govt.nz/courses/find-out-about-study-and-training-options/qualifications-and-their-levels/

GeoNet. (n.d.). https://www.geonet.org.nz/

Health New Zealand. (2023). *Cancer Web Tool*. https://tewhatuora.shinyapps.io/cancer-web-tool/

LAWA. (2023). *Air Quality.* https://www.lawa.org.nz/explore-data/air-quality/

LAWA. (n.d.) *LAWA is a collaboration of organisations with a common aim: to tell the story of our environment.* https://www.lawa.org.nz/about

Minister of Health. (2021). *Methodology.* https://minhealthnz.shinyapps.io/nz-health-survey-2017-20-regional-update/_w_27e6298c/#!/methodology

Statistics New Zealand. *About Us.* https://www.stats.govt.nz/about-us/