FIT5120 Industry experience Studio Project - Onboarding Project

Data Management Plan

- Team: T##.

Contents

1. Datasets overview

To address Victorian parents' concerns about sun protection, the open datasets "Dataset with Cancer Incidence and Mortality by State and Territory from 1982 to 2019", "Location Data", and "UV Index and Temperature" will be utilised to gain insights and knowledge against the parameters related to sun protection.

Data sources					
Names	Physical access (e.g. API, CSV)	Frequency of source updates	Frequency of iteration system updates	Granularity	Copyright/ licensing details
Cancer Incidence from 1982 to 2023 in Australia	CSV	Annually	Annually	High	https://creativecommons.org/licenses/by/3.0/au/ Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
Cancer Mortality from 1982 to 2023 in Australia	CSV	Annually	Annually	High	https://creativecommons.org/licenses/by/3.0/au/ Attribution — You must give

					appropriate credit, provide a link to the licence, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
Location Data	CSV	After System Upgrade	Never	High	GNU General Public License v3.0
UV Index and Temperature	API	Real-time	Minutes	High	https://creativecommons.org/licenses/by-sa/4.0/ You are free to: Share — copy and redistribute the material in any medium or format Adapt — remix, transform, and build upon the material for any purpose, even commercially.

1.1 Open Data Sources

1.1.1 Dataset with Cancer Incidence by State and Territory from 1982 to 2019 (CSV)

https://www.aihw.gov.au/getmedia/e8779760-1b3c-4c2e-a6c2-b0a8d764c66b/AIHW-CAN-122-CDiA-2021-Book-1a-Cancer-incidence-age-standardised-rates-5-year-age-groups.xlsx.aspx

1.1.2 Dataset with Cancer Mortality by State and Territory from 1982 to 2019 (CSV)

https://www.aihw.gov.au/getmedia/9f5cdd1c-87f7-4f05-9a4f-8c5141a3e17e/AIHW-CAN-122-CDiA-2021-Book-2a-Cancer-mortality-age-standardised-rates-5-year-age-groups.xlsx.aspx

1.1.3 Location Data(CSV)

https://gist.github.com/randomecho/5020859

1.1.4 UV index and Temperature(API)

https://openweathermap.org/api/one-call-api

2. Iteration 1

2.1 Data usage

In our project, the datasets play a crucial role in understanding and addressing the challenges posed by skin cancer incidence and mortality rates across various regions. The "Dataset with Cancer Incidence by State and Territory from 1982 to 2019" (CSV) and

"Dataset with Cancer Mortality by State and Territory from 1982 to 2019" (CSV) offer historical perspectives on the prevalence and impact of skin cancer over almost four decades. These datasets enable users to visualise trends, identify age group with higher risks, and formulate targeted interventions and prevention strategies.

The "Location Data" (CSV) combined with real-time "UV Index and Temperature" (API) provide invaluable insights into current environmental conditions. By integrating location-specific information on UV index and temperature, users can receive personalised recommendations for sun protection measures tailored to their immediate surroundings. This data empowers users to make informed decisions about outdoor activities and adjust sun protection measures according to the prevailing environmental factors.

Furthermore, the availability of hourly and daily forecasts based on UV index and temperature data allows users to anticipate periods of heightened UV exposure and extreme temperatures. This proactive approach enables individuals to plan their outdoor activities more safely and take necessary precautions to mitigate the risk of skin damage and related health issues.

2.2 Data Preparation

2.2.1 Data Cleaning and Wrangling

Sunscreen Usage Table Creation

Created the Sunscreen Usage table using SQL query

Cleaning

No cleaning was required for this table as it was a very simple dataset with just 5 rows showing general requirements for sunscreen usage and which SPF to use

Wrangling

Wrangled the dataset to split the UV-Index column into 2 parts Lower bound and Upper bound for easier manipulation.

Mortality and Incidence

- Data Loading: We utilised the readxl package in R to load the datasets named "incidence.xlsx" and "mortality.xlsx". These datasets contain information regarding cancer incidence and mortality, respectively, spanning multiple years and demographic categories.

- Filtering by Cancer Type: We filtered the datasets to include only records related to skin cancer. This step was crucial to focus our analysis on the specific type of cancer we aimed to study.

- Column Renaming: We renamed specific columns to improve clarity and consistency across datasets. For instance, we renamed the "Age-standardised rate(WHO) (per 100,000)" column to "incidence_rate(per 100,000)" in the incidence dataset, and similarly for the mortality dataset.

- Column Selection: We selected only the columns relevant to our analysis, which include "Cancer group/site", "Year", "Age group (years)", "Sex", "Count", and "Incidence/Mortality Rate".

- Data Type Conversion: We ensured that numeric columns such as counts and incidence/mortality rates were stored as numeric data types for accurate analysis. We also rounded the incidence and mortality rates to one decimal place to maintain consistency and precision.

- Filtering by Gender and Year: We further filtered the datasets to include records for males and females only. Additionally, for the mortality dataset, we retained data from 1982 to match with incidence dataset.

- Data Integration: We merged the filtered incidence and mortality datasets based on common attributes such as "Year", "Sex", "Age group (years)", and "Cancer group/site". This integration facilitated a comprehensive analysis of skin cancer trends over time and across demographic categories.

- Data Export: Finally, we exported the merged dataset into a CSV file named "incidence_mortality.csv", ensuring compatibility to import to mysql and ease of access for further analysis and visualisation.

2.3 Data storage

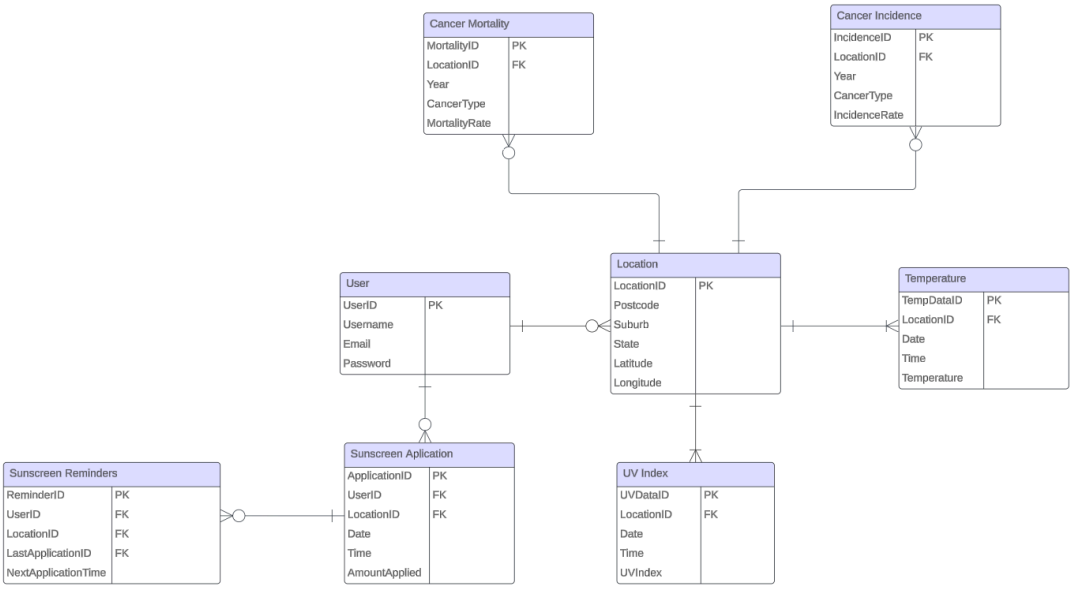
In our data storage strategy, we employ distinct approaches for handling two types of data sources: CSV files and API data. For CSV files, which encompass historical datasets on cancer incidence and mortality rates, we opt for a relational database management system (RDBMS) leveraging MySQL. RDBMS offers a structured framework to organise CSV data

into tables, facilitating efficient storage, retrieval, and manipulation of relational data. On the other hand, for real-time data obtained via APIs, such as UV index and temperature information, we implement a caching mechanism. Caching allows us to store API responses temporarily, minimising latency and optimising data retrieval performance.

2.4 Database design

ER diagram

A representation of the relationship between the data elements is provided by the ER diagram. The structure of the database can be better understood with its assistance.



2.5 Data Analytics

2.5.1 Hindsight Insights

- Cancer Incidence and Mortality Data:
  Historical data on cancer incidence and mortality rates allow users to understand the severity of the skin cancer problem in various regions over time. By analysing past trends, users can identify areas with higher risks of skin cancer and adjust their sun protection measures accordingly.

2.5.2 Foresight Insights:

- UV Index and Temperature Data:
  Real-time UV index and temperature data provide users with insights into the current environmental conditions, enabling them to plan outdoor activities more safely.Hourly and daily forecasts allow users to anticipate periods of heightened UV exposure and extreme temperatures, helping them make informed decisions about sun protection.

2.5.3 Insights:

- Location-Based Recommendations:
  Location data combined with UV index and temperature information allows users to receive personalised recommendations for sun protection measures based on their current location.Users can select appropriate clothing and sunscreen application strategies tailored to the UV index and temperature of their specific location.

- Sunscreen Usage Recommendations:
  Calculations based on UV index data help users determine the optimal amount of sunscreen to apply, ensuring adequate protection against harmful UV rays.By understanding the relationship between UV index levels and sunscreen effectiveness, users can effectively mitigate the risk of sunburn and skin damage.

- Sunscreen Reminders and Tracking:
  Sunscreen reminder features assist users in maintaining consistent sun protection throughout the day, reducing the risk of sunburn and long-term skin damage. Tracking sunscreen applications allows users to monitor their sun protection habits and make necessary adjustments to ensure their children's safety during outdoor activities.