**Assignment 2 Proposal**

**Research Question and Themes (Q1)**

The research question of our group is “what are the connections between urban greening in Victoria and life quality of the residents?”. The research on which this report is based is to inspect the connections between urban greening in Victoria and the life quality for residents. Urban greening in Victoria may affect livability for Victorians in many ways, including cancer rate, suicide rate, mortality rate and also age density in each suburb. Natural environments have been found to have direct influences on human behaviors, thus, areas with different proportions of afforestation could result in different rates, affecting the livability of Victorians.

**Motivations and Investigation Values (Q2)**

The motivation of tackling these questions is because it is anticipated that the findings in this research would provide evidence for potential influences on humans by the environment, hence the result may lead to future changes to increase afforested areas, benefiting the society. The target audience would be households with elders and newborns, environmentalists and government. All Victorians could be affected by the findings of this question. This project might provide evidence between urban greening and suicide rate, cancer rate and age density, assisting the government to further improve resident’s physical and mental health.

**Open Datasets and Explanations (Q3)**

There are four possible open datasets that we could use.

The first dataset is Population estimates by age and sex, by SA2, 2019. <https://www.abs.gov.au/statistics/people/population/regional-population-age-and-sex/2019/32350ds0001_2019.xls>

It is formatted as an Excel spreadsheet.

The second dataset is Melbourne Metropolitan Region Urban Vegetation Cover: 2018.

<https://www.planning.vic.gov.au/__data/assets/pdf_file/0018/440172/CompiledReport_MelbourneMetro_v2.1.pdf>

It is formatted as a PDF file.

The third dataset is CANCER IN VICTORIA Statistics & Trends 2019.

<https://www.cancervic.org.au/downloads/cec/cancer-in-vic/Cancer-in-Victoria-statistics-and-trends-2019.pdf>

It is formatted as a PDF file.

The last dataset is about the Suicide rate.

<https://www.aihw.gov.au/getmedia/aca3fd17-a3dc-4ec8-b6d8-be8e62a46b12/aihw-suicide-and-self-harm-monitoring-nmd-suicide-icd-10-x60-x84-y87-0.xlsx.aspx>

It is formatted as an Excel spreadsheet.

Also, this dataset is about the mortality rate of every state in Australia.

<https://www.aihw.gov.au/reports/cancer/cancer-data-in-australia/contents/cancer-by-state-and-territory-data-visualisation>

It is an html website, which displays the data in graph and table form.

Datasets are collected from many resources, including Australian Bureau of Statistics, Cancer Council Victoria, Planning Victoria and etc. These datasets include Victorian age density, percentage of Tree coverage, cancer statistics and suicide rates. The size of these files varies between kbs to mbs, containing raw dataset for the rates mentioned above.

Dataset 2 and 4 can be categorized by suburbs and try to observe relationship between suicide rate and percentage of greening in each area.

Dataset 1 and 4 can be categorized by suburbs, intending to compare the population proportion and age density with the percentage of green areas in each area.

Dataset 3 and 4 will also be categorized suburbs, seeking to find relationships between cancer rates and Greening areas.

**Analysis Techniques and Investigation Methodologies (Q4 & 5)**

Data wrangling techniques would pre-process the collected data into corresponding order that best provides a clear table result. Different ways of organizing and formatting are convenient to analyze the relationships between datasets, dataframe and groupby functions may be used to achieve this.

Using data wrangling methodologies, data is arranged into different graphs, histogram (age), line chart (suburbs) and pie chart(rates) and tables of data. Raw dataset contains information that is not discovered yet. Pre-processing and connecting data in the way that links them together could provide relationships between the subjects, which later can be analyzed clearly and efficiently.

**Challenges and Risks (Q6)**

Challenges

Firstly, one of the challenges might be data collecting. Since the collected dataset cannot match in exact year, the dataset can be unreliable if it is collected a few years ago, or with some data that are missing or wrong. If there is insufficient collected data, then the connection between urban greening and Victorian’s livability could be invalid.

Secondly, links between the datasets may also be one of the challenges, as each dataset could require different process skills, and relationships between the dataset could be hard to discover.

Moreover, one of the open datasets that we find is a html file in the website. It only displays the data it has in its own database, but does not provide easy access to visitors. Then, it is difficult for us to analyse by using the csv style.

Risks

•Timeliness: without a specified date or not effective any more

•Lack of resources to complete the report

•The primary dataset is collected with incorrect cells.

•Dataset does not cover enough related or effective information.