# Introduction

This document comprises the information, usage, and all the details of the data sources that are used by the GOT-G team to develop the GirlSafe application and website. The process of how the data being updated or transposed into the application, including the analysis or modelling is also described here.

1. data governance

data tools and techniques  
privacy  
retention  
ethics  
security

# 2. Data Sources (Open Datasets) Details

Here are the detailed information of the data sources used by the GOT-G team:

| **Names** | **Physical access** | **Frequency of source updates** | **Frequency of iteration system updates** | **Granularity** | **Copyright/licensing details** |
| --- | --- | --- | --- | --- | --- |
| Crime records in Victoria  Link: <https://www.crimestatistics.vic.gov.au/crime-statistics/latest-victorian-crime-data/download-data> | CSV downloaded | Quarterly ( every 4 months) | 2 month after source updates | Number of crime records recorded in each suburb in Victoria | <https://creativecommons.org/licenses/by/4.0/>  When reporting this data you must attribute the Crime Statistics Agency (or CSA) as the source. |
| Suburb/ Locality Boundaries in Victoria  Link:  <https://data.gov.au/data/dataset/vic-suburb-locality-boundaries-geoscape-administrative-boundaries/resource/02dd5644-7542-4e95-ab71-1707010b99a6> | Shapefile downloaded | Quarterly | 2 month after source updates | Digital boundaries (spatial data) of the suburbs in Victoria | <https://creativecommons.org/licenses/by/4.0/>  Attribution for Licensed Material:  Administrative Boundaries © Geoscape Australia licensed by the Commonwealth of Australia under Creative Commons Attribution 4.0 International licence (CC BY 4.0). |
| Australian Postcode  Link:  <https://github.com/Elkfox/Australian-Postcode-Data/blob/master/au_postcodes.csv> | CSV downloaded | Last updated on December 23, 2016 | - | Postcode, latitude, and longitude of each suburb in Australia | [https://creativecommons.org/licenses/by/3.0/](https://creativecommons.org/licenses/by/4.0/)  The original source for this material is found at <https://github.com/Elkfox/Australian-Postcode-Data/blob/master/au_postcodes.csv> The dataset has been reformatted to fit the general purpose of the application |
| Feature Lighting  Link:  <https://data.melbourne.vic.gov.au/City-Council/Feature-Lighting-including-light-type-wattage-and-/4j42-79hg/data> | CSV downloaded | Weekly (last updated October 2, 2021) | - | Location, lighting type and wattage of feature lighting across the City of Melbourne. | Data will be made available under flexible and open licenses, allowing for reuse by the public.  <https://creativecommons.org/licenses/by/4.0/> |

Other open datasets that can be explored (Backup):

<https://www.abs.gov.au/articles/sexual-violence-victimisation#key-statistics>

<https://www.abs.gov.au/statistics/people/crime-and-justice/personal-safety-australia/2016#data-download>

<https://www.aihw.gov.au/reports/domestic-violence/sexual-assault-in-australia/data>

<https://www.abs.gov.au/statistics/people/crime-and-justice/recorded-crime-victims/latest-release#data-download>

# 3. Dataset Usage

## Epic 2: Dangerous Area

We use the criminal records from the CSA (Crime Statistics Agency) to show the crime rate that can be used to indicate how safe the suburb is. The records are then visualised in a map so that the user can easily get an insight into the information. The suburb name and the postcode are used to connect the two datasets and can be used as the value for the user to filter the dataset based on their desire.

# 4. Data wrangling and cleaning

Pre-processing steps are performed with Python to clean the datasets:

1. Crime records dataset
   * The dataset is downloaded from the CSA site.
   * Check on duplicates or missing data in the dataset to make sure the dataset is clean and reliable.
   * Filter the dataset, only use the criminal records of the "Crimes against the person".
   * Adjust (remove the sub-title) the 'Offence Subdivision'
   * Remove unnecessary columns.
   * Export clean dataset to CSV format file.
2. Digital suburb boundaries
   * The shapefile dataset is downloaded from the source link.
   * Check on duplicates or missing data in the dataset to make sure the dataset is clean and reliable.
   * Remove unnecessary columns.
3. Australian Postcode Dataset
   * The dataset is downloaded from the source link.
   * Check on duplicates or missing data in the dataset to make sure the dataset is clean and reliable.
   * Filter the dataset for Victoria areas only.
   * Remove unneeded columns.
   * Export clean dataset to CSV format file.

Further Data processing

* Combine the postcode and suburb boundaries dataset.
* Impute missing values.
* Impute missing latitude and longitude with the centroid of the area.
* Generate the location\_id based on the suburb name and postcode.
* Generate the tables (location and crime) with the location id as the relation between the two tables.
* Import the data into the database
* Perform data exploration and provide the result on the website for user insight.

# 5. db details Dataset Storage

The modified crime dataset is stored in the database (local) while the map dataset is stored in the drive with CSV file format (called using a link) since the size is too big and takes too much memory for the application.



Figure 1: ER diagram plan for iteration 1

Link to sql script: <https://drive.google.com/drive/folders/1pCIvtYZ6yg5oSaquO2R-Usu0PCKbpxah>

# Data Security

This work is licensed under the [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/). You may share and redistribute the material in any medium or format. When using the material you must attribute the GOT-G as the source.

# User Privacy

This application may collect the information from the user such as the location, phone number of the closest person (emergency contact) to be utilised in the application features. The data is confidential and will not be shared with other users or other parties without the permission from the user. There will be a message prompted to let the user know that the application needs their permission to collect their personal information.

# 6. Data Reflection

## Iteration 1

The second Epic, which is the "Dangerous Area" can be fulfilled with the criminal records dataset, digital boundaries shapefile, and the postcode dataset. Using these datasets and the data visualisation, we can accomplish one of the goals which is to provide some insight to the user about the safe and unsafe areas in Victoria so they can raise their awareness.

After the implementation of the dataset to develop the features, there is an ambiguity within the limits to determine the safe and unsafe areas. There are some areas with no criminal records (either missing data or no criminal action reported) which need further exploration and checking. The population of each suburb can also be one of the factors for the high records of criminal records that happen. Thus, we need to take this into consideration to provide a piece of reliable information for the user.

Collecting the open dataset of the population in each suburb and providing the crime rate based on the population might be more reliable rather than the number of criminal records that happen in each suburb. We can use the population to provide the probability of a crime that might happen in a particular area for the user to take a concern to.