FIT5120 Industry Experience Studio Project

Data Governance and Management

****

**TA25 GOT-G:**

Yang Ge

Wenjun Ma

Regina Valentina

Xinyu Wang

Yining Zhou

# Table of Contents

[**Table of Contents**](#_z3p61p8leifh) **2**

[**Introduction**](#_368d53sijsbj) **3**

[**Data Sources (Open Datasets) Details**](#_af637msyzpvy) **3**

[Dataset Details](#_bohamv4cx1e) 3

[Data Gathering](#_y2b8ebp9y4lg) 5

[Crime records in Victoria (area)](#_rnr29rlhp0vq) 5

[Suburb/ Locality Boundaries in Victoria](#_mg0tiej4wyyk) 6

[Australian Postcode](#_rdosn9racn3k) 6

[Feature Lighting](#_4ur0sf8yzdl) 7

[Crime records in Victoria (Victim report)](#_822msn6y2o86) 8

[**Dataset Usage**](#_o1ksfl803f4g) **8**

[Epic 2: Dangerous Area](#_kg4t065zwnh2) 8

[Epic 7: Victoria Crime Records Visualisation](#_bfa6zuayfntt) 9

[**Data wrangling and cleaning**](#_y7ehfr8b7eo8) **9**

[Crime records dataset](#_lmy51xckshcm) 9

[Digital suburb boundaries](#_t7cptfr31los) 12

[Australian Postcode Dataset](#_leckt4lfq59c) 13

[Further Data processing](#_pjryzhkom19u) 15

[**Database Details Dataset Storage**](#_di0v3xy5mhum) **19**

[**Data Security**](#_y2yzeqz2qaip) **20**

[**User Privacy**](#_68brwtgidwhe) **20**

[**Data Reflection**](#_wg0l3g99gtsy) **20**

[Iteration 1](#_5tk3qipclznm) 20

[Data Insight](#_kwgvoycb3c) 20

[Data Hindsight](#_5nx8bhi5ix3t) 20

[Data Foresight](#_7uhw4lfb3wbs) 21

[Iteration 2](#_ypjuvo1m4gkx) 21

[Data Insight](#_2xpnaua7867p) 21

[Data Hindsight](#_m2p4twdnm898) 21

[Data Foresight](#_pjy59u5y5s7j) 21

[Iteration 3](#_w1b14je32zv0) 21

[Data Insight](#_3luow6jxqs1f) 21

[Data Hindsight](#_kbxs37xkad2c) 22

[Data Foresight](#_cr3ntbvmycqp) 22

# **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.

# **Data Sources (Open Datasets) Details**

## Dataset 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 (area)  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 |
| Crime records in Victoria (Victim report)  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 Local Government Area in Victoria with the Victim detail (age range and gender) | <https://creativecommons.org/licenses/by/4.0/>  When reporting this data you must attribute the Crime Statistics Agency (or CSA) as the source. |
| 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>

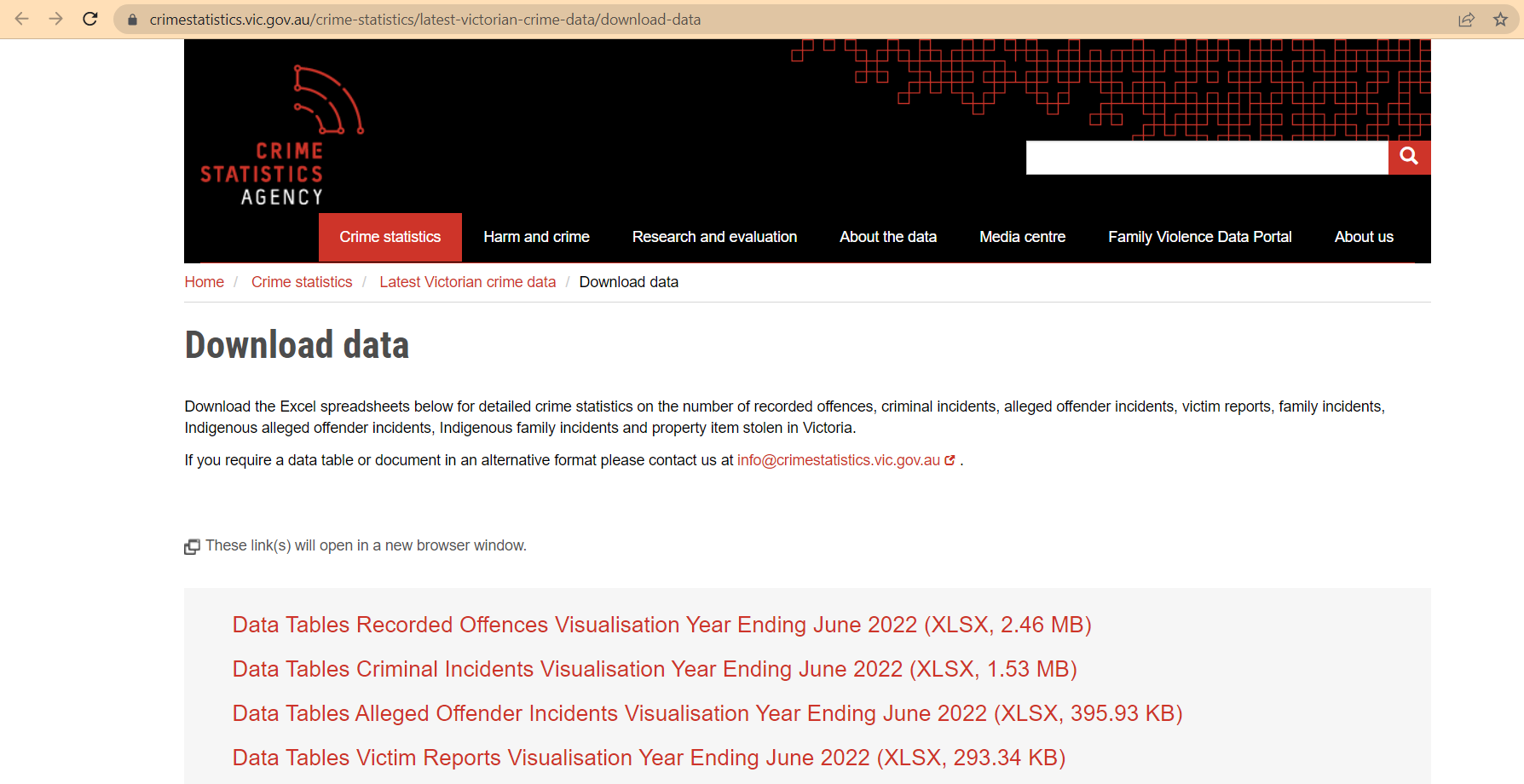
<https://www.abs.gov.au/statistics/people/crime-and-justice/recorded-crime-victims/latest-release#data-download>

## Data Gathering

In this section we provide the step-by-step of how to collect the dataset from the link described in the previous section.

### Crime records in Victoria (area)

* Go to the given website link



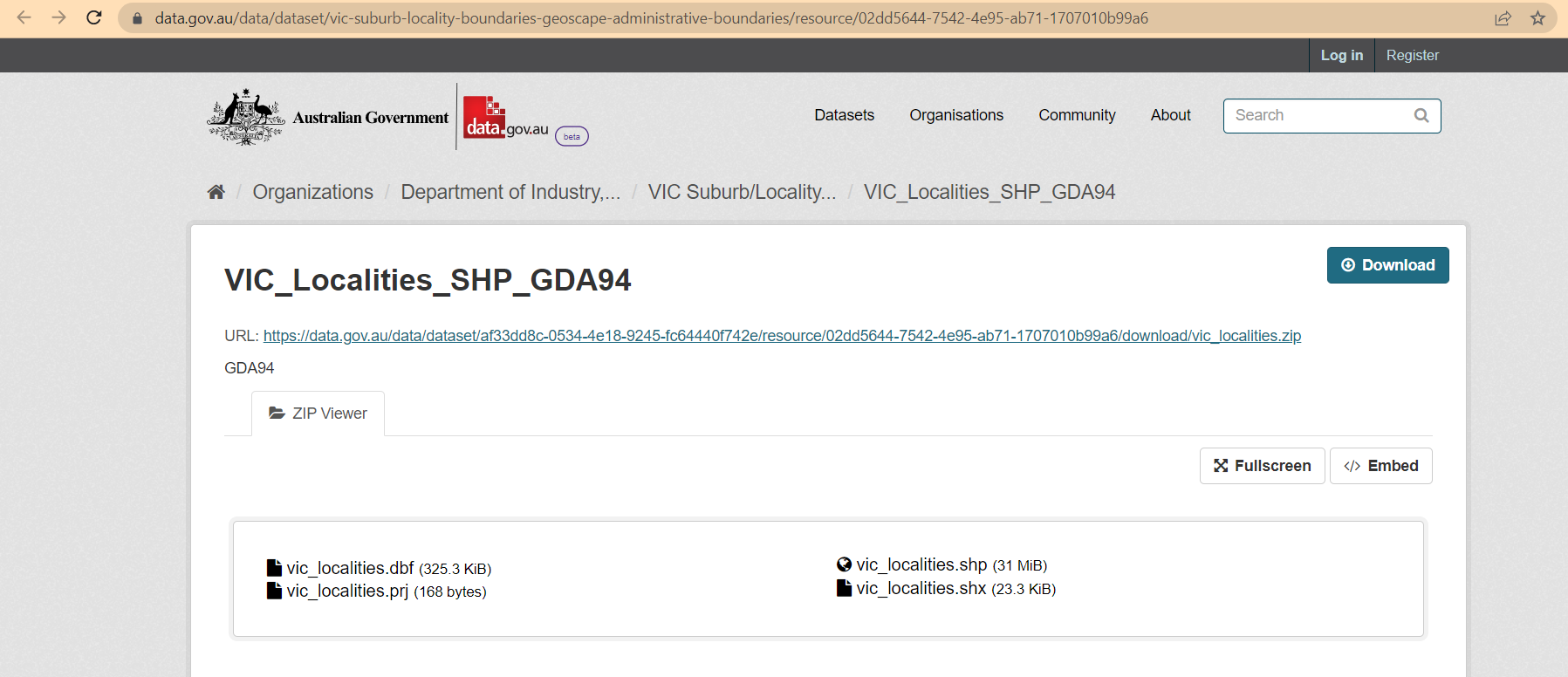
* Scroll down and find the “Data\_Tables\_LGA\_Criminal\_Incidents\_Year\_Ending\_March\_2022.xlsx”



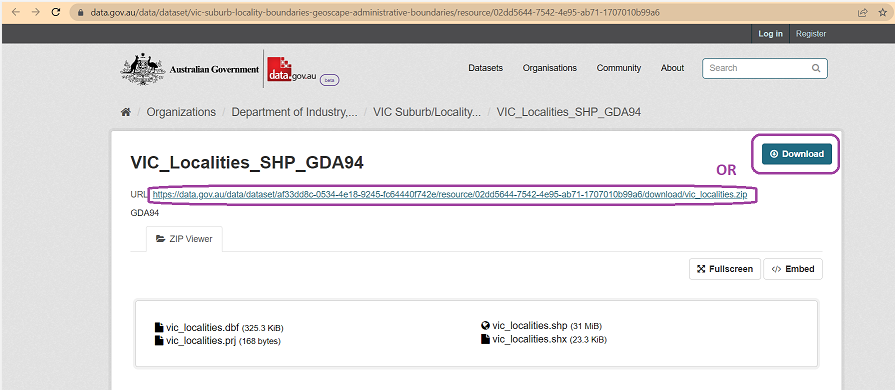
* Click on the text to automatically download the dataset

### Suburb/ Locality Boundaries in Victoria

* Go to the website link as described previously.



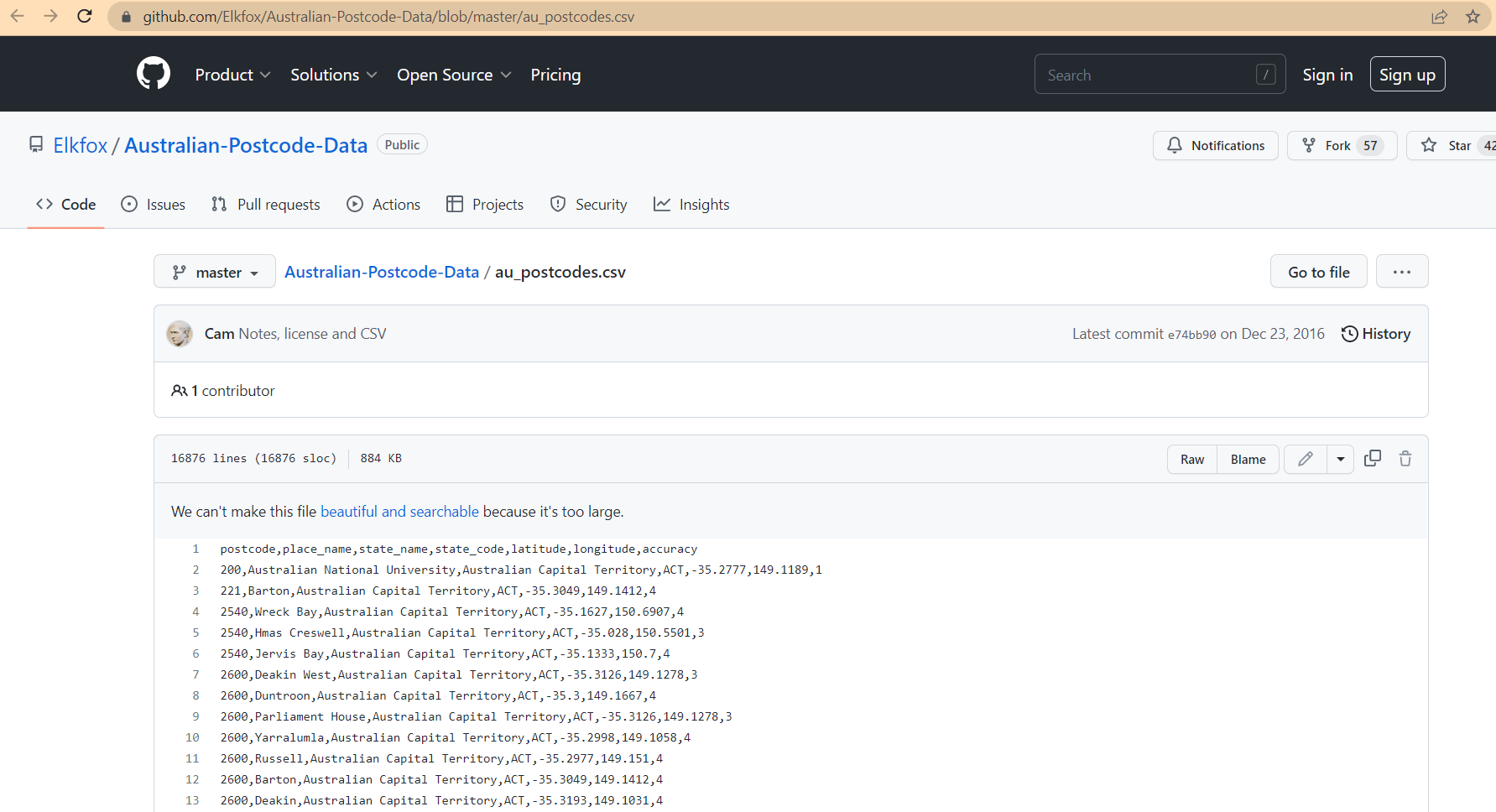
* Click either the “download” button or the given URL to download the dataset.



* The dataset will be automatically downloaded in a `zip` file format.

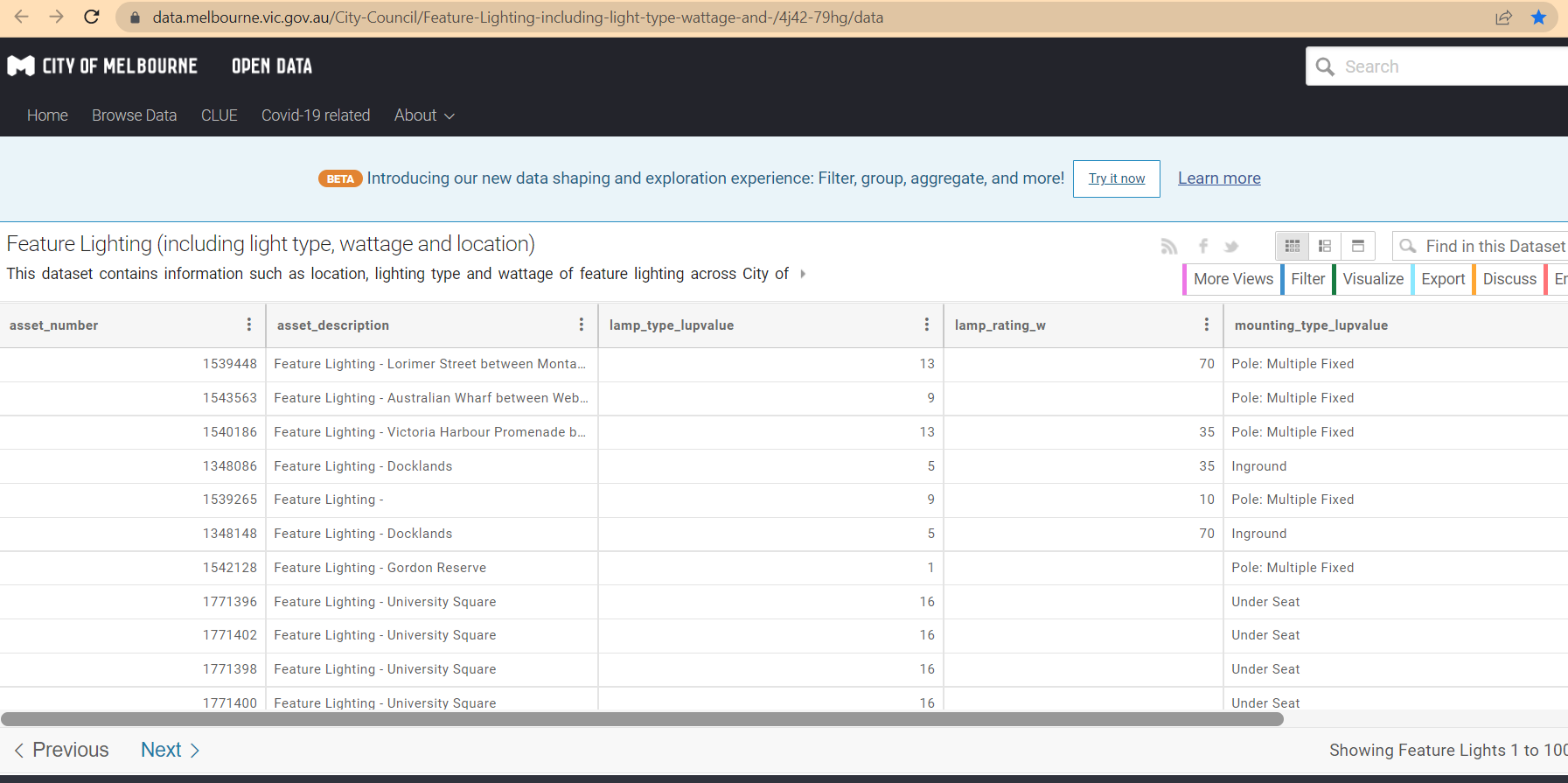
### Australian Postcode

* Go to the described website link.

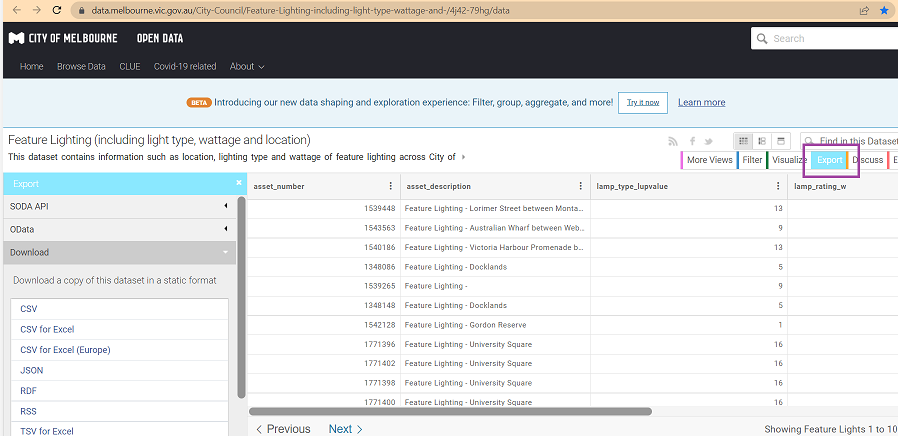


### Feature Lighting

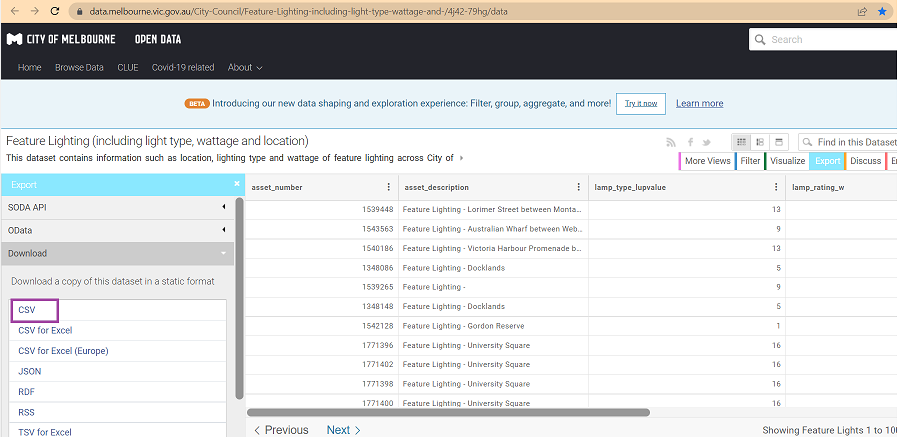
* Go to the website link as described.



* Click on the “export” button.



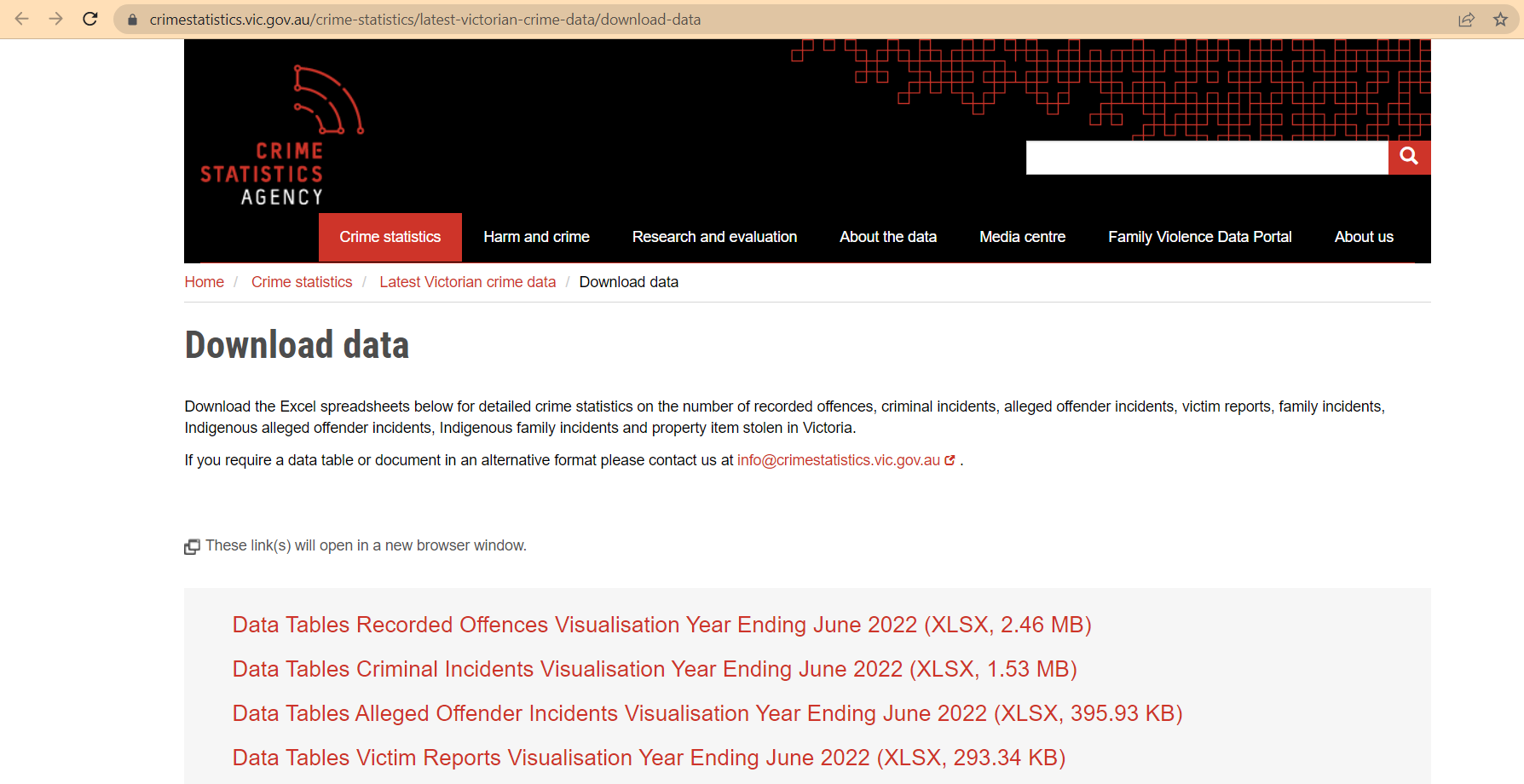
* Save the dataset in the desired file format (in this project, we use the CSV format).



* The dataset will automatically downloaded into the local computer

### Crime records in Victoria (Victim report)

* Go to the given website link



* Scroll down and find the “Data\_Tables\_LGA\_Criminal\_Incidents\_Year\_Ending\_March\_2022.xlsx”



* Click on the text to automatically download the dataset

# Dataset Usage

## Epic 2: Dangerous Area

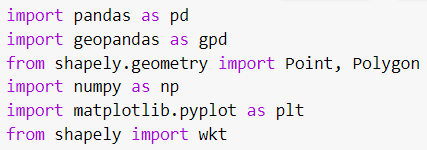
We use the criminal records from the CSA (Crime Statistics Agency) to show the crime rate that can be used to indicate how safe the suburb is. The records are then visualised in a map so that the user can easily get an insight into the information. The suburb name and the postcode are used to connect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

## Epic 7: Victoria Crime Records Visualisation

As we get the dataset of the crime records that happen in Victoria, we utilise the dataset to gain a piece of information that can help the user to be more aware. Here, we tried to provide the summary of the dataset and get the essential points for the user such as the crime rate that happens in Victoria, the type of crime that exists in Victoria, and the number of street harassment that are recorded in each suburb in Victoria.

# Data wrangling and cleaning

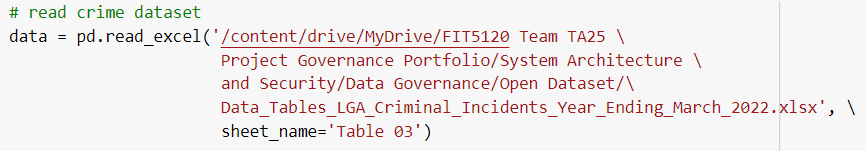
The data pre-processing steps are performed using Google colab with Python. We also upload the open datasets downloaded to google drive for the access purpose. Some of the libraries are also used to process the data.



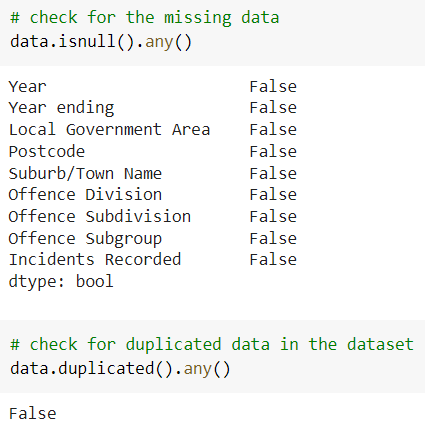
Furthermore, here are the steps performed to check and clean the obtained open datasets:

## Crime records dataset

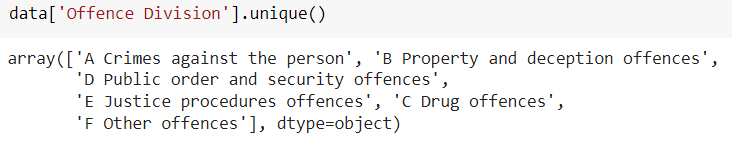
* The dataset is downloaded from the CSA site.
* Upload the dataset to the related google drive.
* Read the data



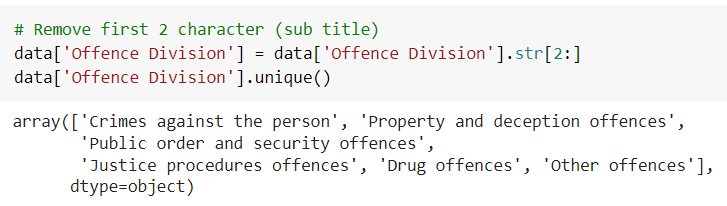
* Check on duplicates or missing data in the dataset to make sure the dataset is clean and reliable.



* Get the Offence Division of each crime records



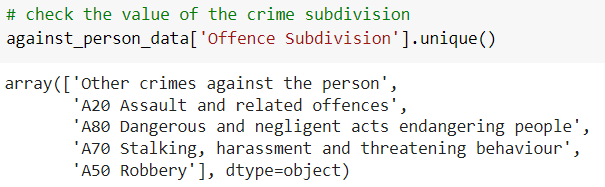
* Remove the Heading title of the Offence division



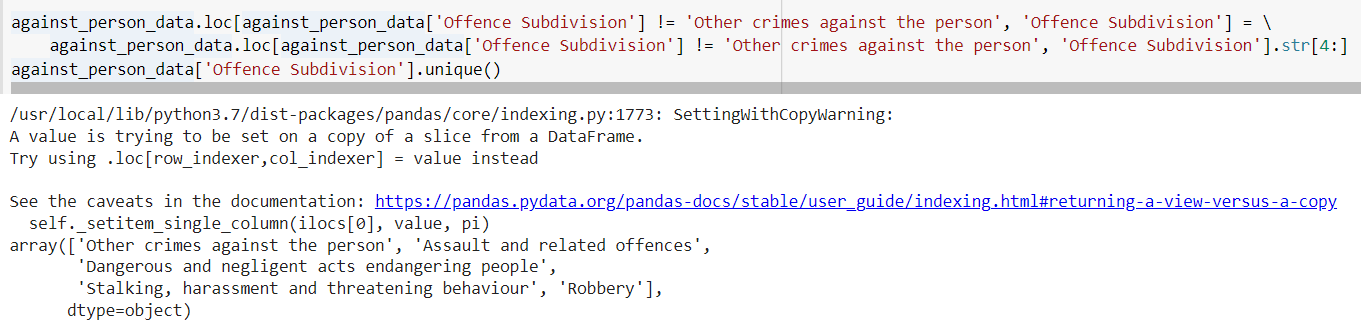
* Filter the dataset, only use the criminal records of the "Crimes against the person".



* Check the modification made



* Remove the Heading title of the Offence subdivision, but exclude the crime type 'Other crimes against the person'



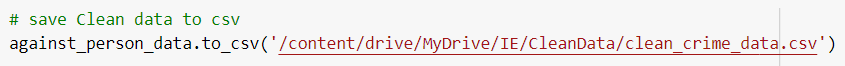
* Adjust (remove the sub-title) the 'Offence Subdivision'
* Remove unnecessary columns



* Group and sum the crime records from the Offence Subgroup

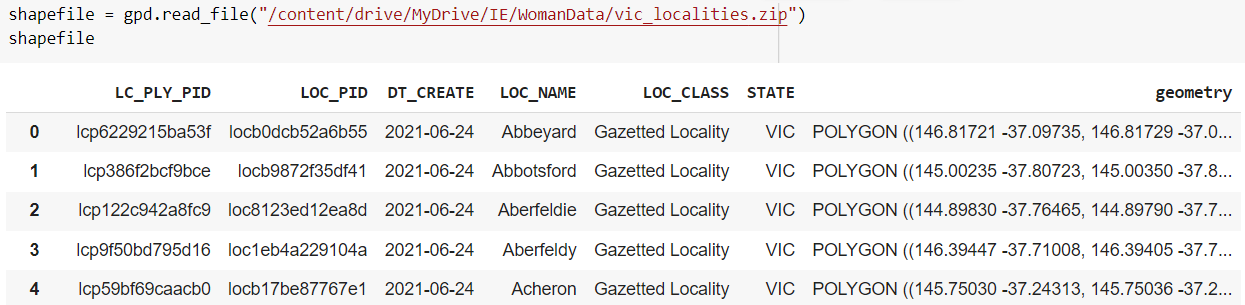


* Export clean dataset to CSV format file.

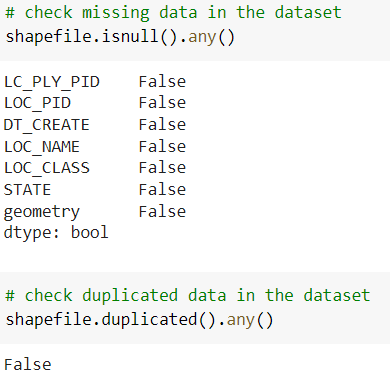


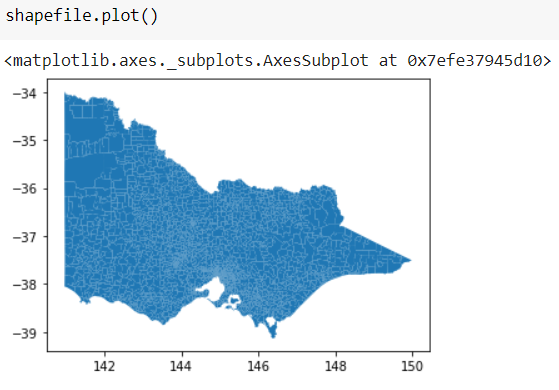
## Digital suburb boundaries

* The shapefile dataset is downloaded from the source link.
* Upload the file to the google drive for access.
* Read the shapefile dataset in the zip file format.

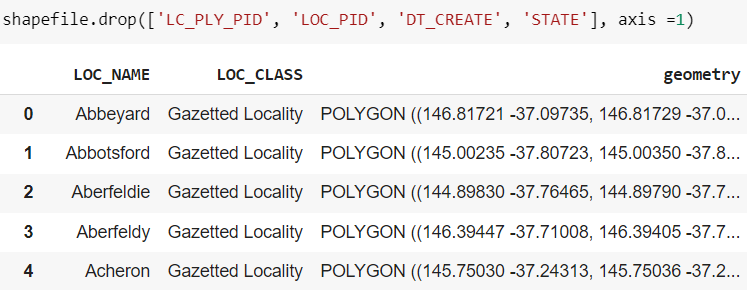


* Check on duplicates or missing data in the dataset to make sure the dataset is clean and reliable.





* Remove unnecessary columns

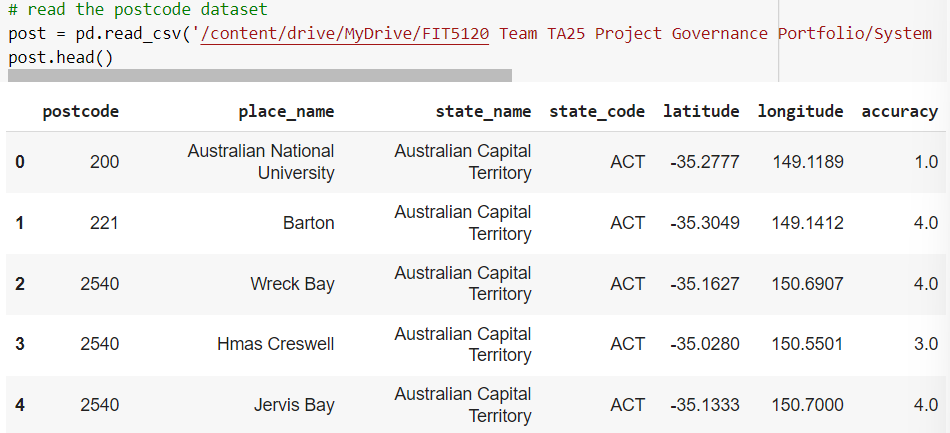


* Export the dataset

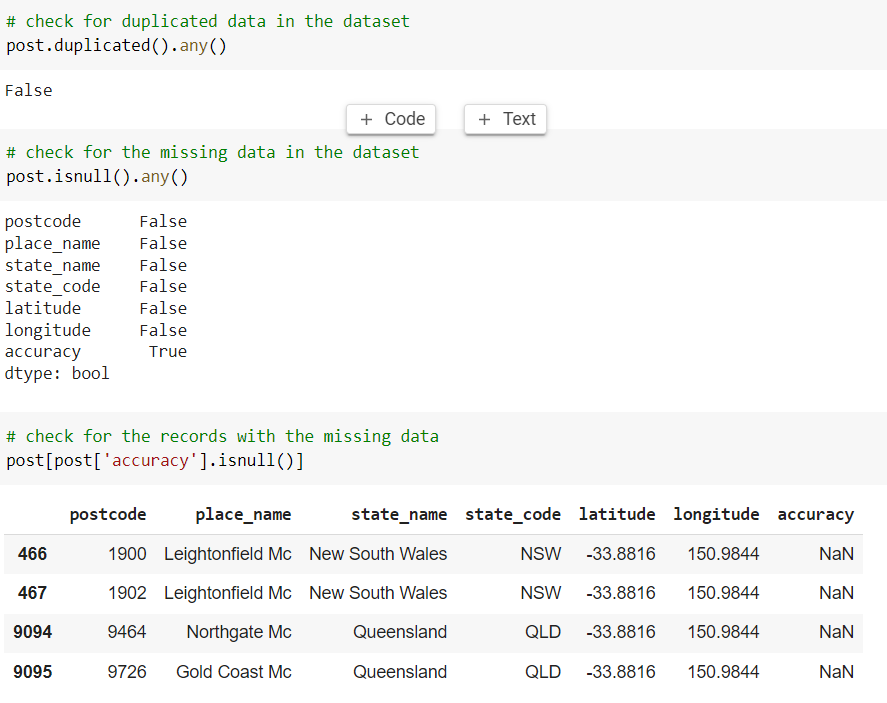


## Australian Postcode Dataset

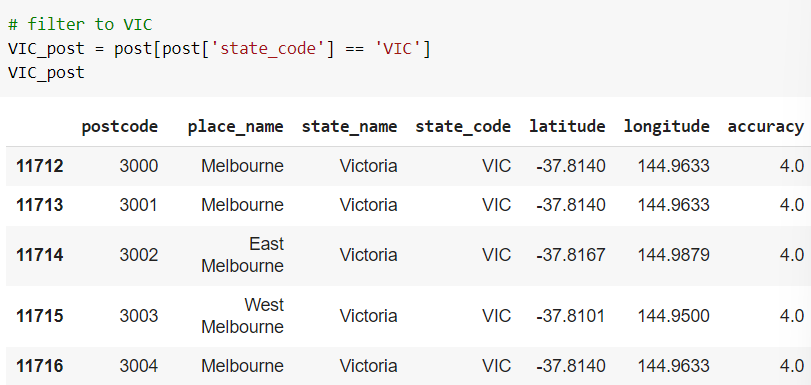
* The dataset is downloaded from the source link.
* Upload the dataset to the drive for access in the google colab.
* Read the excel formatted dataset.



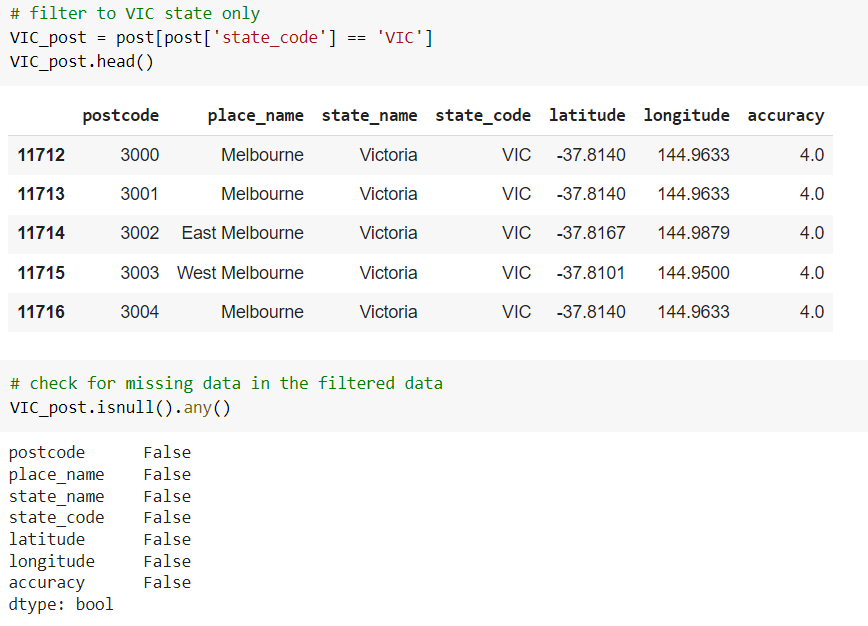
* Check on duplicates or missing data in the dataset to make sure the dataset is clean and reliable.



* Filter the dataset for Victoria areas only.



* Re-check the missing data in the filtered dataset.



* Remove unneeded columns, including accuracy which contain a lot of missing values.

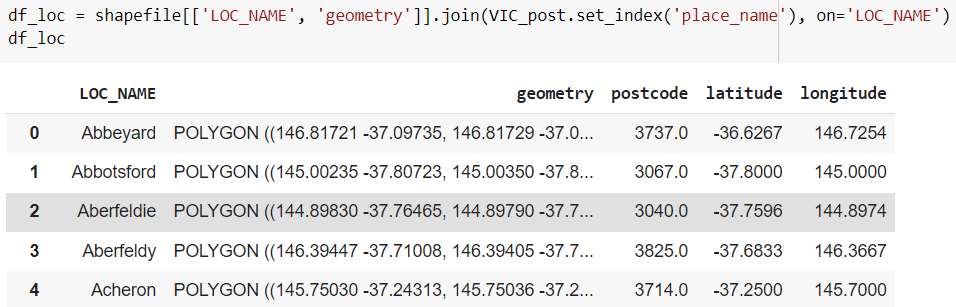


* Export clean dataset to CSV format file.

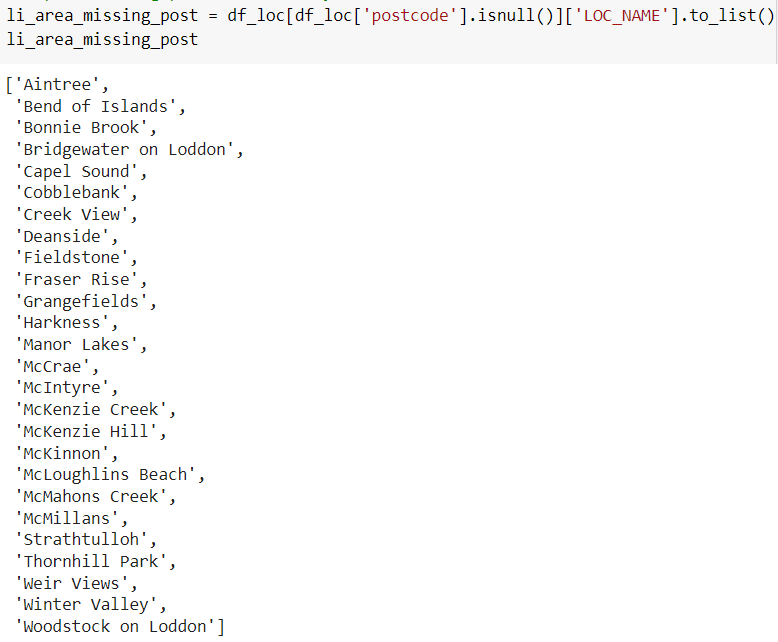


## Further Data processing

* Join the postcode and suburb boundaries dataset.

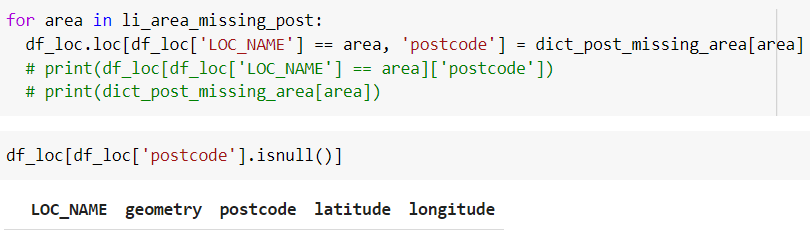


* Check for the area with missing values.

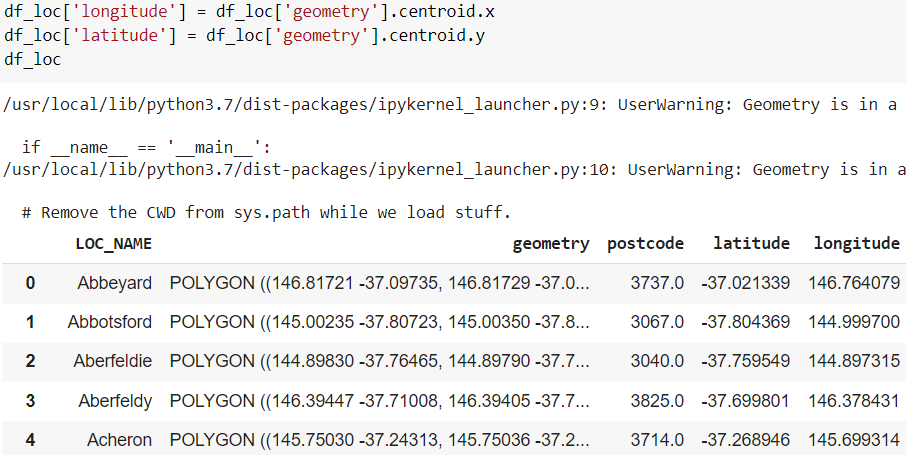


* Impute missing values.

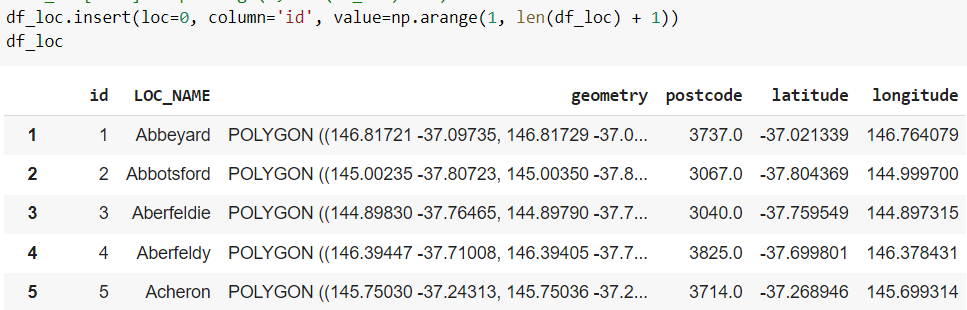


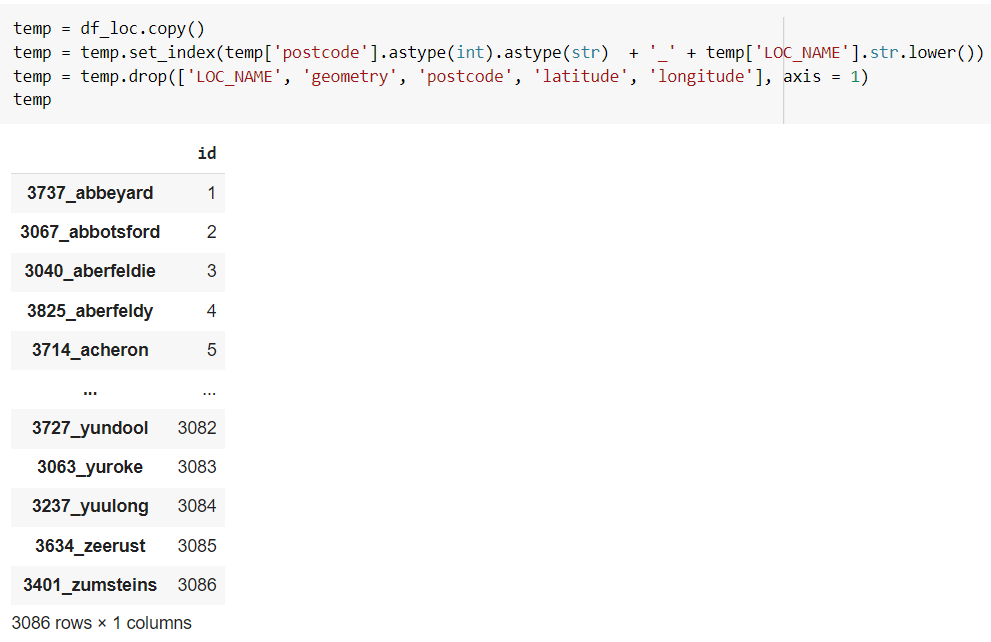


* Impute missing latitude and longitude with the centroid of the area.

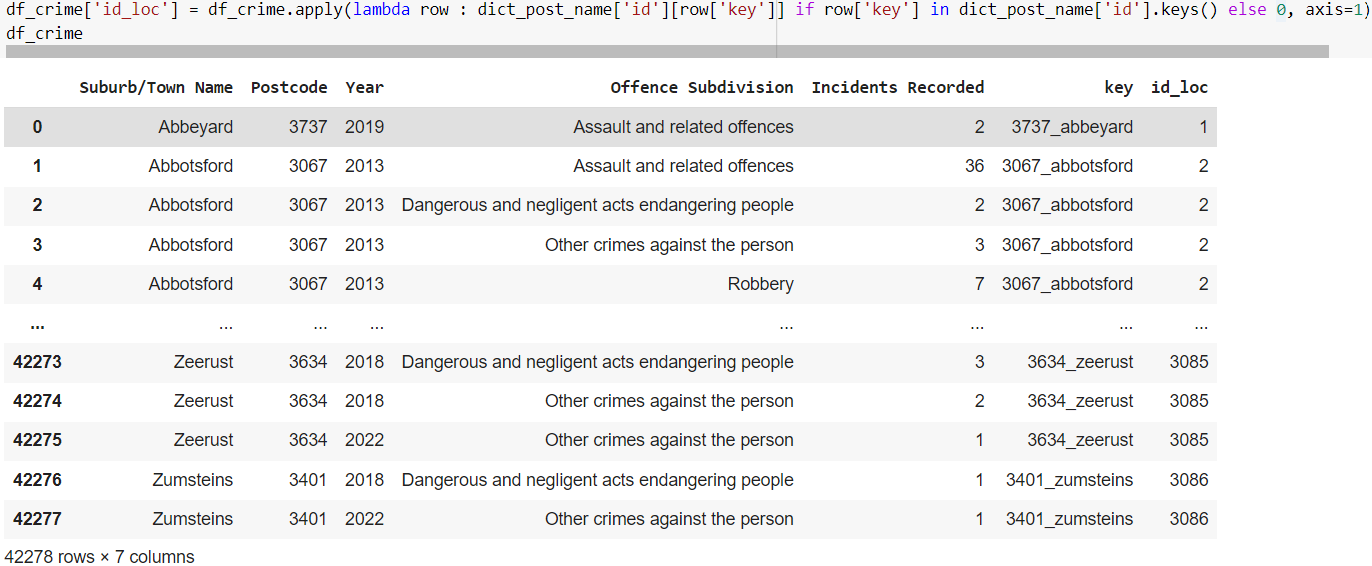


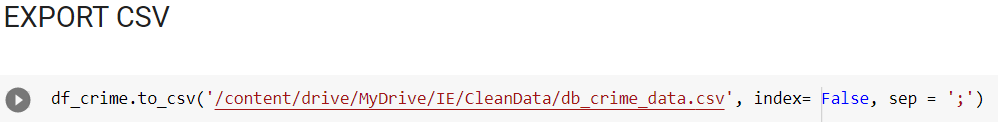
* Add id column for the suburb dataset



* Generate the location\_id based on the suburb name and postcode key as reference to the location id
* 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.

The code for the data preparation and preprocessing can be accessed from this [link](https://colab.research.google.com/drive/16UpaezzZLrgtc-fWATUDcz6MrMw3eghn#scrollTo=MtyhE3ULrntC).

# Database 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.

For further information related to the security aspects related to this application can be found in the [Security Folder](https://drive.google.com/drive/folders/1Z1KohWrmH4eouEixDBk78qdpojRWfXED).

# 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.

# Data Reflection

## Iteration 1

### Data Insight

The seventh Epic, which is the "Victoria Crime Records Visualisation" is achieved using the criminal records dataset. Using the dataset, we provide the user a simple data visualisation to show the crime trend that happened in Victoria for the past 10 years. As we look back on what we did for the first iteration, the way we presented the data to the user is not friendly and engaging. As we only provide the dataset in the form of a barchart, line chart, and a simple table, the user might not be able to obtain the information we want to deliver from the data.

### Data Hindsight

This kind of decision is one of the common mistakes that happen when we perform a data visualisation, where we only focus on visualising the data and neglect the user’s point of view. Whereas the user’s perspective is supposed to be the most important thing that we consider in the first stage of the data visualisation.

### Data Foresight

In the future iteration, we need to adjust and fix this issue by taking the user’s point of view as one of the consideration factors to visualise the data. Secondly, we can also deliver the important pieces of information as a story along with infographic images to obtain the user’s attention and make the information easier to understand. Therefore, the data can be fully utilised to provide a piece of reliable information and understanding to raise the user awareness about the topic we have.

## Iteration 2

### Data Insight

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.

### Data Hindsight

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.

### Data Foresight

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.

## Iteration 3

### Data Insight

To provide more information about the safety to the user, we make use of the Crime records in Victoria based on the victims reports. This dataset has a similar structure with the previous crime dataset used (crime dataset in Victoria based on the area). Therefore, it is easier to perform the dataset checking and cleaning to gain the information from the data. The difference from the previous dataset is that this dataset provides some information about the victim’s age, gender, etc.

### Data Hindsight

Because of the short amount of time given, we cannot fully explore the dataset and gain the most information out of it. Thus, we restrict the dataset to obtain the relation between the crime that happens in Victoria with the victim’s age and gender.

### Data Foresight

For the future works, we can explore the dataset from different factors. Moreover, there are still a lot of other dataset that we can explore to get the information related to the safety in Victoria. Therefore, in the future, we think it would be good if we can use or combine more open datasets to get the fullest information out of it.