图片包含 徽标

描述已自动生成

**DATA1002 Group Assignment Stage 1**

**Group**

**470000946 Xiaobing Wang**

**470021332 Shiyu Hu**

图形用户界面, 文本, 应用程序, 信件

描述已自动生成

# Section 1 Description of data sets

The aim of this report is to investigate the number of Covid-19 confirmed cases in different local government areas. Other attributes that can be used to group by are the population density, the proportion of the elderly, and month of the date.

We selected data from three different sources. The dataset is obtained from official and trustworthy source such as ABS and NSW data.

## 1.1 data sources and the rights associated with the data

Raw Dataset 1， Population estimates by Local Government Area, 2018 to 2019, accessed from <https://www.abs.gov.au/statistics/people/population/regional-population/latest-release#data-downloads-geopackages>

Raw Dataset 2， Population estimates by age and sex - summary statistics, by LGA, 2019, accessed from <https://www.abs.gov.au/statistics/people/population/regional-population-age-and-sex/latest-release#data-download>

Raw Dataset 3, NSW COVID-19 tests by location and result, accessed from

<https://data.nsw.gov.au/data/dataset/covid-19-cases-by-location>

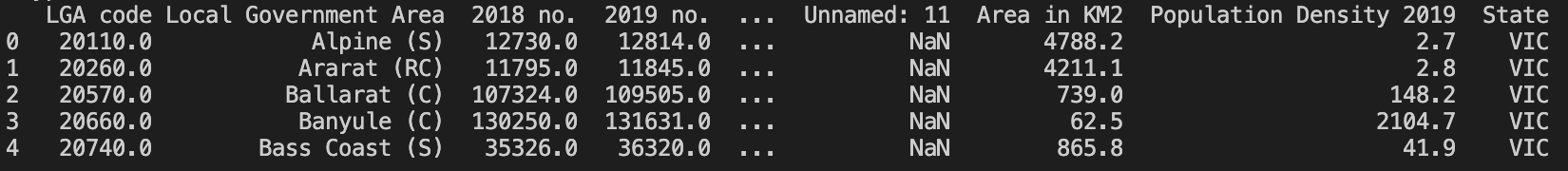
Both dataset 1 and dataset 2 authorized by The Australian Bureau of Statistics, a famous official Statistics organization of Australia. Dataset 3 is authorized by Data.NSW, a famous official Statistics organization supported by NSW government. These data are open to free use and gives us trusted official statistics for comprehensive research and use of these data follows the [NSW Government Open Data Policy](https://data.nsw.gov.au/sites/default/files/inline-files/NSW_Government_Open_Data_Policy_2016.pdf) defined in the Government Information (Public Access) Act 2009 (NSW) (GIPA Act).

## the format/contents of the data

|  |  |  |
| --- | --- | --- |
| Attribute | Description | Type and form |
| LGA code | Local Government Area code | String – digits |
| LGA Government Area | Local Government Area Name | String |
| 2018 no. | 2018 population number | String – digits |
| 2019 no. | 2019 population number | String – digits |
| Population Change | Population Change between 2018&2019 | String – digits |
| Change in % | Population Change in % | String – digits |
| Natural Increase | Population Natural Increase | String – digits |
| Net Internal Migration | Population Net Internal Migration | String – digits |
| Net Overseas migration | Population Net Overseas migration | String – digits |
| Area in KM2 | Local Government Area in km^ 2 | String – digits |
| Population Density 2019 | Population Density 2019 | String – digits |
| State | State name in short form | String |

* D1\_population\_by\_lga.csv

This dataset contains estimates of the resident population of Local Government Areas of Australia for 30 June 2018 and 30 June 2019, according to the 2019 edition of the Australian Statistical Geography Standard (ASGS). Estimates are revised for 2018 and preliminary for 2019.

This dataset contains 8565 values and provides population density of various Local Government Areas in Australia. The formats of the data are shown as below, together with the first 5 rows displayed.

* D2\_Age\_Summary\_by\_lga.csv

This dataset contains preliminary estimates of the resident population by age and sex as at 30 June 2019. Data are provided for Local Government Areas (LGAs) of Australia, according to the 2019 edition of the Australian Statistical Geography Standard (ASGS). This dataset contains 8190 values and provides age summary of various Local Government Areas in Australia.

The formats of the data are shown as below, together with the first 5 rows displayed.

文本

描述已自动生成

|  |  |  |
| --- | --- | --- |
| Attribute | Description | Type and form |
| S/T code | Suburb & Town | String – digits |
| S/T name | Suburb & Town Name | String |
| LGA code | Local Government Area code | String – digits |
| LGA Name | Local Government Area Name | String |
| Males no. | Males number | String – digits |
| Females no. | Females number | String – digits |
| Persons no. | Total number | String – digits |
| Sex ratio males per 100 females | Sex ratio males per 100 females | String – digits |
| Median age in years | Median age in years | String – digits |
| People aged 0-14 years in % | People age between 0-14 number in total | String – digits |
| People aged 15-64 years in % | People aged between 15-64 number in total | String – digits |
| People aged 65 years and over in % | People aged 65 years and over in total | String - digits |

* D3\_confirmed\_cases\_by\_lga\_postcode.csv

This dataset is about COVID-19 confirmed cases in New South Wales – COVID-19 data including notification date and postcode, local health district, and local government area. The dataset is updated daily, except on weekends.

This dataset contains 25182 values and provides age summary of various Local Government Areas in Australia. The formats of the data are shown as below, together with the first 5 rows displayed.

文本

描述已自动生成

|  |  |  |
| --- | --- | --- |
| Attribute | Description | Type and form |
| notification\_date | The date of notification | String – year/month/day |
| postcode | Suburb & Town postcode | String - dights |
| lhd\_2010\_code | Local health district code | String - digits |
| lhd\_2010\_name | Local health district name | String |
| lga\_code19 | Local Government Area code | String - digits |
| lga\_name19 | Local Government Area name | String |

## 1.3 some comments on any strengths or limitations of the dataset

The official datasets are accurate and useful for research, adding credibility to the conclusion of the study. However, come of the test cases are having unknown sources, limiting the researchers to fully investigate every confirm cases and group by local government areas. In some research, these unknown cases are ignored and rows of these unknown cases are excluded to make the number of cases grouped by area more reliable and sensible.

Due to time constraint, the datasets does not include more factors that may be important in influencing number of cases in certain area, such as access to health service, number of commuters and medical care awareness. With more database to further improve the depth of this project, more meaningful insights can be drawn to help the general public to know more about the pandemic.

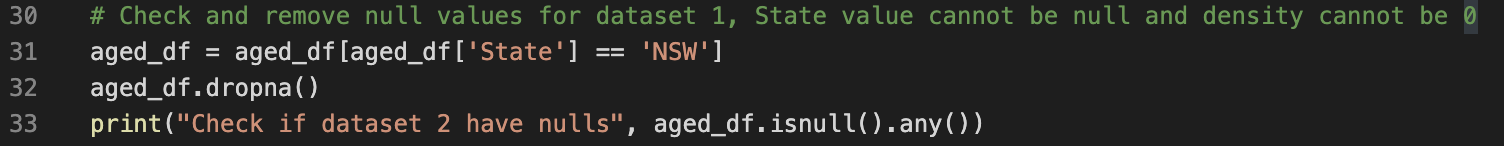
# Section 2 Data transformation

### 2.1 Check and remove null values, fill in missing values

Code are used to check and remove null values for datasets and ensure there is no null values for dataset 1 and 2

文本

描述已自动生成



Code are used to remove null values for dataset 3. NA values are filled as ‘Unknown’ to represent confirm cases with unknown sources and code is used to ensure there is no null values for dataset 3.

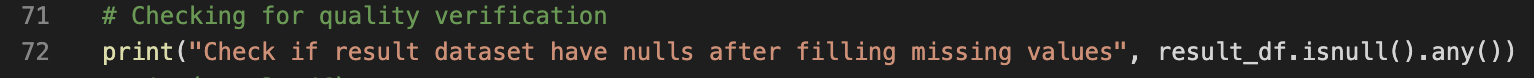
文本

描述已自动生成

Code are used to replace NA values with meaningful strings. Rows with numeric attributes as NaN are removed. Code is used to explicitly show that no null value is present.

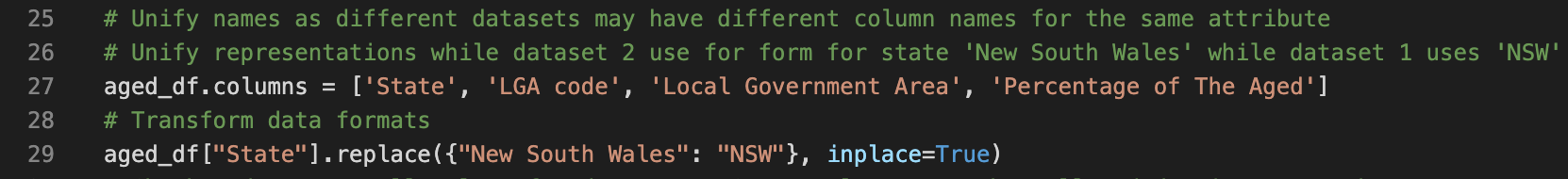
文本

描述已自动生成

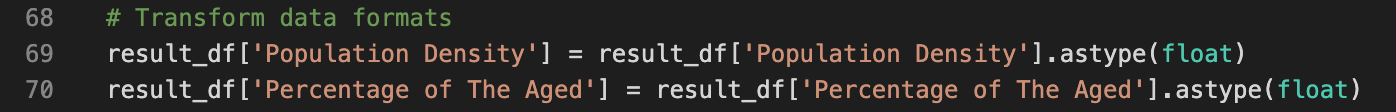


### 2.2 Transform data formats

Datasets with different column names for the same attribute are being modified to have consistent column names. While dataset 2 use for form for state 'New South Wales' while dataset 1 uses 'NSW', dataset 2 is modified to be consistent with dataset 1.

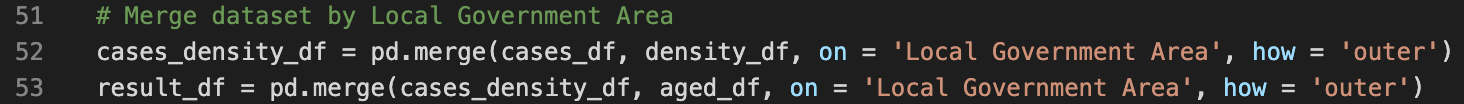


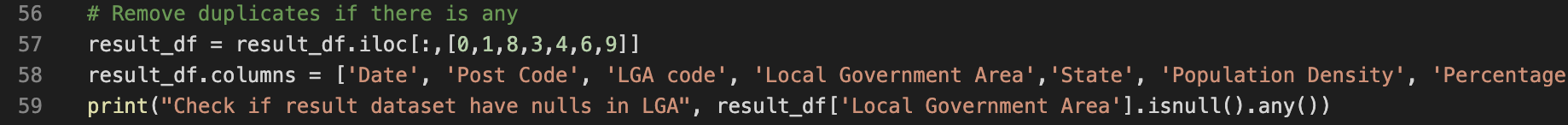
Numeric attributes are converted to floats from string of digits before being saved into the result dataset.



### 2.3 Merge dataset and remove duplicates

Datasets are merged using the attribute the Local Government Area. This attribute is checked to be consistent in all datasets. Duplicates are removed.





2.5 Automated checking for quality verification

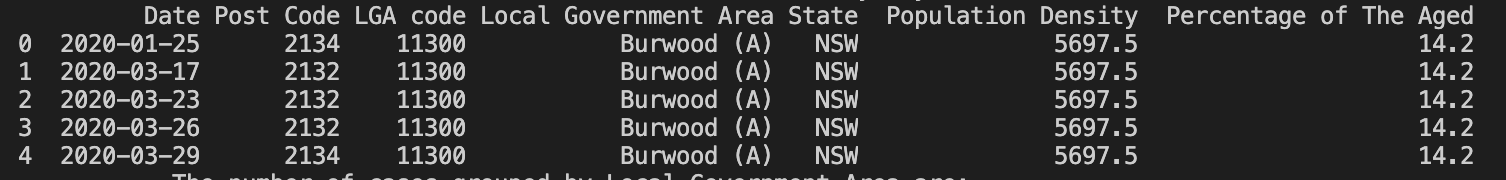
Beside checking done after each dataset is cleaned and transformed. The final clean dataset is checked again to ensure there is no null values and all values are meaningful for further processing.

文本

描述已自动生成

2.6 Metadata of the merged dataset (这个应该和clean data 在一起)

This clean dataset provides confirmed cases of various Local Government Areas in Australia, together with population density, percentage of the elderly statistics. The formats of the data are shown as below, together with the first 5 rows displayed.



|  |  |  |
| --- | --- | --- |
| Attribute | Description | Type and form |
| Date | The date of notification | String – year/month/day |
| postcode | Suburb & Town postcode | String of digits |
| LGA code | Local Government Area code | String of digits |
| LGA Government Area | Local Government Area Name | String |
| State | State in short form | String |
| Population Density | Population divided by area | float |
| Percentage of The Aged | Percentage of population over 65 | float |

# Section 3 Analysis of the data

## 3.1 number of cases, grouped by Local Government Area

The result is analysed and the number of cases are grouped by Local Government Area.

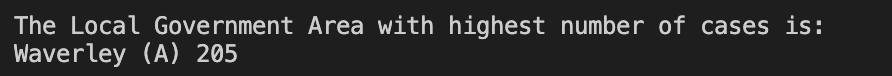
文本

描述已自动生成

文本

描述已自动生成

The highest number of cases is shown after grouping counts by Local Government Area, excluding unknown cases.

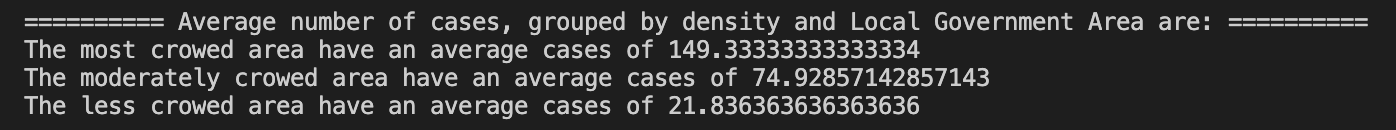


## 3.2 Average number of cases, Grouped by population density and Local Government Area

Population Density is categorized into 3 groups, excluding unknown cases. Cases are calcaluated and grouped by population density. The averages number of cases of each group are shown.

文本

描述已自动生成



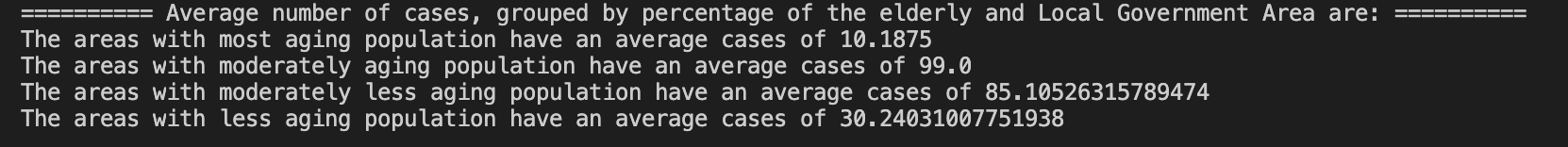
## 3.3 Average number of cases, are Grouped by percentage of the olderly

Local Government Area that needs special care using Number of cases and aging population percentage

Aging percentage is spitted into 3 groups, excluding unknown cases. The averages number of cases of each group are shown.

文本

描述已自动生成



3.4 Number of cases, Grouped by month

文本

描述已自动生成

文本

描述已自动生成

Highest number is also shown.

