Project 2

Team 3 Technical Document

The aim of this project was to develop a data resource that enables analysis of COVID cases by postcode and Local Government Area (LGA) compared to the population and population density of each LGA.

We collected data and cleaned it using Pandas. Our Data Frames were cleaned using Pandas Drop and Filter Functions.

We also used the Python Lambda function to split Information from one column and create two new columns. We also used the DataFrame.rename to get correct column names.

Duplicates were removed using Df.drop_duplicates function.

Data Sources:

The three data sets sourced in this project are:

1. Commonwealth Australian Bureau of Statistics (ABS), population by LGA

2. LGA Mapping file

3. The Victorian DHHS COVID data:

1. Commonwealth Australian Bureau of Statistics (ABS), population by LGA

(a) data source - url:"https://api.data.abs.gov.au/data/C21_G04_LGA/....2"

(b) Format: Data set sourced via API

(c) Extract and Load method: Sourced via abs API

(d) Original data fields: DATAFLOW, AGEINGP: Age, SEXP: Sex, REGION: Region, REGION_TYPE: Region Type, STATE: State, TIME_PERIOD: Time Period, OBS_VALUE

(e) Data transformation:

(i) REGION: Region: separate data in REGION: Region column into 'Regions ID' and 'Region Name'

(ii) Drop DATAFLOW, AGEINGP: Age, REGION: Region, STATE: State

(iii) SEXP: Sex filter and retain 3: persons value in SEXP: Sex column - this removed the male and female categories and enabled to calculation of the total population

(v)Rename 'OBS_VALUE' to 'Population'

(vi) Drop 'persons' column to avoid duplicating data

(f) Comments

This data set could not be directly merged with the Victorian DHHS COVID data set. The 'Region Name' field in this data set did not include suffix (x) that was present in the ‘Localgovernmentarea' field.

(g) Final data fields: Population, Region ID (PRIMARY KEY) and Region Name

2. LGA Mapping and LGA SQKM file

(a) data source: Resource/'LGA_2020_VIC2.csv'

(b) Format: CSV

(c) Extract and load method: data downloaded from Vic Government website

(d) Original Data Fields: MB_CODE_2016, LGA_CODE_2020, LGA_NAME_2020, STATE_CODE_2016, STATE_NAME_2016, AREA_ALBERS_SQKM

(e) Data transformation

(i) Drop MB_CODE_2016, STATE_CODE_2016, STATE_NAME_2016

(f) Comments

This data set was required to map the LGAs in the abs and COVID data sets. This data set also provides the area square kilometer data to enable population density for each LGA to be shown

(g) Final data fields: LGA_CODE_2020 int (PRIMARY KEY) LGA_NAME_20201 object (FORIEGN KEY), AREA_ALBERS_SQKM float

1. The Victorian DHHS COVID that shows the number of COVID cases by LGA by date (day, Month and year)

(a) data source: "https://www.dhhs.vic.gov.au/ncov-covid-cases-by-lga-source-csv"

(b) Format: CSV

(c) Extract and Load method: read in directly from the DHHS web site as too large to download and save

(d) Original data fields: diagnosis date, Postcode, acquired, Localgovernmentarea.

(e) Data transformation:

(i) use the 'diagnosis date' to create a column called 'year' to enable the volume of people to be grouped by year

(ii) drop the 'acquired' field

(f) Comments:

This data set could not be directly merged with the abs population data set. The 'Localgovernmentarea' field in this data set did not include the LGA numerical ID and contained an additional code in the form of(x)at the end of the name. This code was not present in the abs data set to enable data to be merged a third data set is required to match the LGA name and ID. This data set did not show population density

(g) Final Data Fields

This resource forms the foundations which, at a future date, can enable the addition and analysis of other socio, demographic economic factors that are captured in Census data and other State and Commonwealth data sets

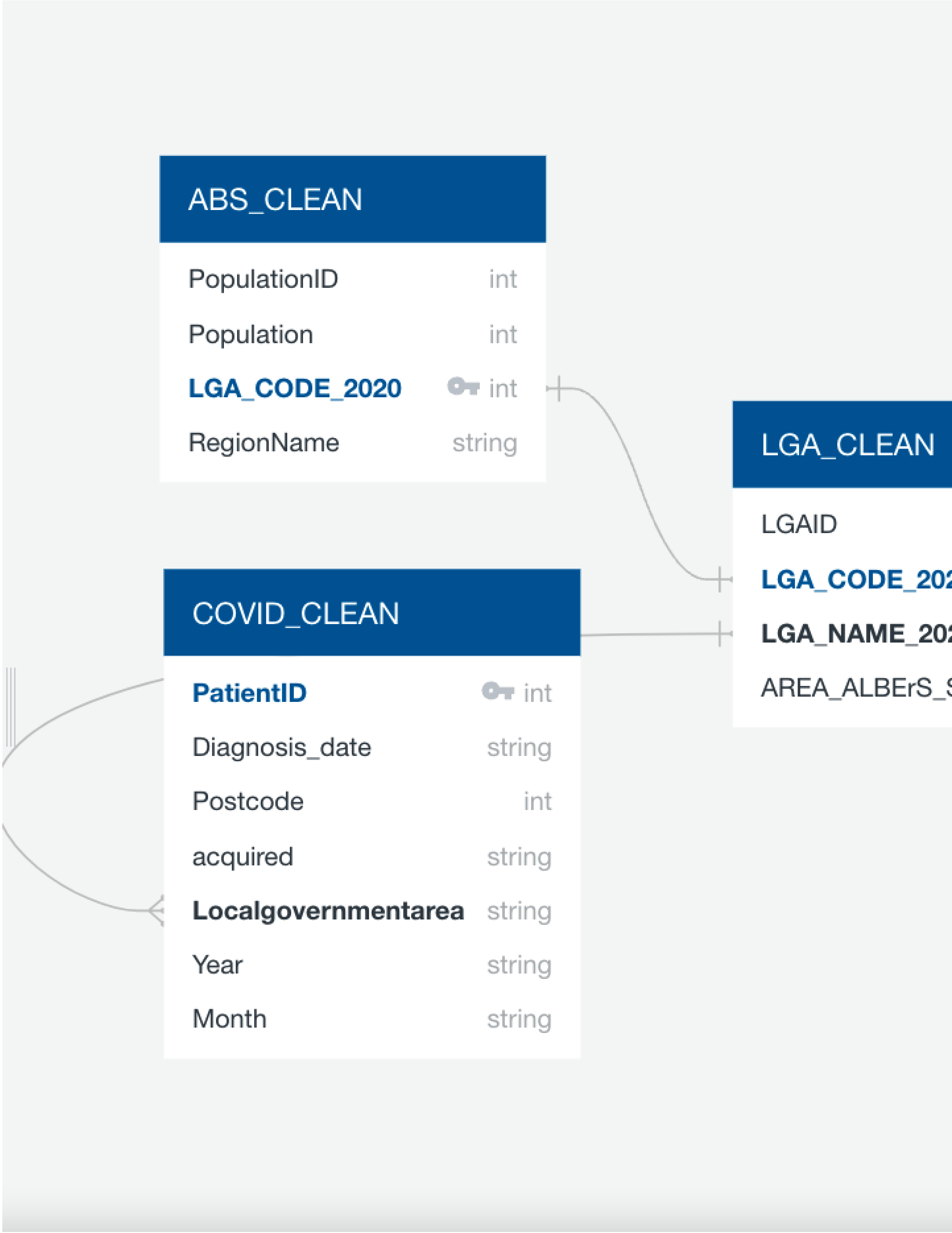
SQL

We created a new database, Project2, using PG Admin. We created 3 tables that reflected the columns and data types from the new CSVs acquired from cleaning data.

We loaded the data into the database using SQL Alchemy by;

a) Storing CSV files into a data frame
b) Connecting to the local database
c) Using pandas to load csv converted Data Frame into database
d) Using pandas to load json converted Data Frame into database

ERD Diagram



ABS_CLEAN		
PopulationID		int
Population		int
LGA_CODE_2020	🔑	int
RegionName		string

COVID_CLEAN		
PatientID	🔑	int
Diagnosis_date		string
Postcode		int
acquired		string
Localgovernmentarea		string
Year		string
Month		string

LGA_CLEAN
LGAID
LGA_CODE_202
LGA_NAME_202
AREA_ALBErS_S