FIT5196-S2-2021 assessment 3

This is an individual assessment and worth 30% of your total mark for FIT5196.

Due date: 6 PM, November 1, 2021

For this assessment, you are required to write Python code to integrate several datasets into one single schema and find and fix possible problems in the data. Input and output of this assessment are shown below:

Table 1. The input and output of the task

Inputs	**Output file**	**Jupyter notebook and .py files**
<student_no>.xml (file), <student_no>.json (file), Vic_suburb_boundary (directory), Vic_GTFS_data (directory), Lga_to_suburb.pdf (file), covidlive.com.au (website) (you need to figure out how to scrap this website by yourself. Check week 9 materials for more info. Any python package is allowed)	<student_no>_A3_solution.csv	<student_no>_ass3.ipynb <student_no>_ass3.py

The .py file should be generated from your jupyter notebook file and it will be used for plagiarism checks.

Each of you is given several datasets in various formats and the initial data is about housing information in Victoria, Australia. **You can find your own dataset here**. In this assignment, you need to perform the following tasks.

Task 1: Data Integration (55%)

In this task, you are required to integrate the input datasets from several sources into one dataset with the following schema.

Table 2. Description of the final schema

COLUMN	DESCRIPTION
property_id	A unique id for the property
lat	The property latitude
lng	The property longitude
addr_street	The property address
suburb (15%)	The property suburb. **Default value: "not available"**
Lga (10%)	The property **local government area** (LGA). **Default value: "not available"**
closest_train_station_id (10%)	The closest train station to the property using Haversine distance. **Default value: 0**
distance_to_closest_train_station (5%)	The Haversine distance from the closest train station to the property. **Default value: 0**
travel_min_to_MC (15%)	So we assumed that there was a big vaccination centre placed at Melbourne Central building. This column is the **rounded average travel time** (minutes) of the **direct journeys** (see the definition of the direct journeys in **Note 2**) from the closest train station to the "Melbourne Central" station on weekdays (i.e. Monday-Friday) **departing** between **7 to 9am**. For example, if there are 3 direct trips departing from the closest train station to the Melbourne Central station on weekdays between 7-9am and each take 6, 7, and 8 minutes respectively, then the value of this column for the property should be round((6+7+8)/3). If **there are no direct journeys** between the closest station and Melbourne Central station, the value should be set to **"not available"**. **If the closest station to a property is Melbourne Central station itself, then the value should be set to 0**. **Default value: -1**
direct_journey_flag (15%)	A Boolean attribute indicating whether there is a direct journey to the Melbourne Central station from the closest station between 7-9am on the weekdays. This flag is 1 if there is a direct trip (i.e. no transfer between trains is required to get from the closest train station to the Melbourne Central station) and 0 otherwise. **Default value: -1**

30_sep_cases (5%)	The number of Covid-19 positive cases on the **30th of September** of the LGA of each property scrapped from covidlive.com.au website. **Default value: "not available"**
last_14_days_cases (5%)	The **rounded average** of Covid-19 cases for the property's LGA for the last 14 days starting from **29th of September backward** (i.e., 29th, 28th, 27th, etc) scrapped from covidlive.com.au website. **Default value: "not available"**
last_30_days_cases (5%)	The rounded average of Covid-19 cases for the property's LGA for the last 30 days starting from 29th of September backward (i.e., 29th, 28th, 27th, etc) scrapped from covidlive.com.au website. **Default value: "not available"**
last_60_days_cases (5%)	The rounded average of Covid-19 cases for the property's LGA for the last 60 days starting from 29th of September backward (i.e., 29th, 28th, 27th, etc) scrapped from covidlive.com.au website. **Default value: "not available"**

Note 1: the output csv file must have the exact same columns as specified on the schema. Please note that the output files which are not in a correct format, as specified in the integrated schema, won't be marked.

Note 2: direct journey means that you can reach the Melbourne Central station without changing your train at any point in the journey. So, when you board the train on the closest station, you can directly go to the Melbourne Central station.

Note 3: if you decide not to calculate any of the required columns, then you must still have that column in your final dataframe with the 'default value' as the value of all the rows. Please note that the output files which are not in a correct format, as specified in the integrated schema, won't be marked.

Note 4: No external data is allowed to calculate the values of the integrated schema. For example, to calculate the suburb, you can only use the shape files provided in the Google drive. The only external source of information is covidlive.com.au website.

Note 5: shapefile data and lga_to_suburb.pdf data can be outdated and incorrect. You don't need to fix them or check their validity.

Note 6: for Haversine distance (link), use 6378 km as the radius of the earth.

Note 7: for more information about GTFS files read here (link).

Note 8: In table 2, the numbers in front of some of the columns in the format of (a%) are the allocated mark associated with that column. For example, column "suburb" carries 15% of the total output mark of task 1. Also, please note that we are aware that the summation of percentages is 90%. The other 10% goes to the issue(s) that may appear during data integration tasks and you should find and resolve them.

Task 2: data reshaping (20%)

In this task, you need to study the effect of different normalization/transformation methods (i.e. standardization, minmax normalization, log, power, box-cox transformation) on the columns scrapped from the covidlive.com.au website (i.e., 30_sep_cases, last_14_days_cases, last_30_days_cases, last_60_days_cases) and observe and explain their effect assuming **we want to develop a linear model to predict the "30_sep_cases" using "last_14_days_cases", "last_30_days_cases", "last_60_days_cases" attributes**. When reshaping the data, we have two main criteria. First, we want our features to be in the same scale and second, we want our features to have as much linear relationship as possible with the target variable (i.e., 30_sep_cases). You need to first explore the data to see if any scaling or transformation is necessary (if yes why? and if not, also why?) and then perform appropriate actions and document your results and observations.

Task 3: Documentation (25%)

The main focus of the documentation would be on the quality of your explanation on task 2 but similar to the previous assignments, your notebook file should be in a decent format with proper sections and subsections.

Deliverables

You must submit the following files on Moodle to have a successful submission. **We will not mark incomplete submissions!**

1. <student_no>_A3_solution.csv
2. <student_no>_ass3.ipynb (which contains both task 1 and task 2 documentation each having their own sections) (having a ToC is highly recommended) **(all outputs must be preserved in the .ipynb)**
3. <student_no>_ass3.py (this file is used for plagiarism checking)