FIT5196-S2-2022 assessment 3

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

Due date: **Wednesday 2nd November**

For this assessment, you are required to write Python (Python 3) code to integrate several datasets into one single schema and find and fix possible problems in the data. The input and output of this assessment are shown below:

Table 1. The input and output of the task

Inputs	**Output**	**Jupyter notebook**
vic_suburb_boundary.zip, gtfs (directory) Crimebylocation.xlsx <student_no>.csv council.txt	<student_no>_solution.csv	<student_no>_ass3.ipynb

You are given multiple datasets in various formats and the task is creating housing information in Victoria, Australia. Your assessment is to perform the following tasks. All the required data can be found here.

Task 1: Data Integration (60%)

In this task, you are required to integrate these datasets into one with the following schema.

Table 2. Description of the final schema

COLUMN	**DESCRIPTION**
ID	A unique id for the property
Address	The property address
Suburb (20/100)	The property suburb. The suburb must only be calculated using Vic_suburb_boundary.zip. **Default value: "not available"**
Price	The property price
Type	The type of property
Date	Date of sold

Rooms	Number of bedrooms
Bathroom	Number of bathrooms
Car	The number of parking spaces of the property
Landsize	The area of the property
Age	The age of the property at the time of selling
Lattitude	The Latitude of the property
Longtitude	The Longitude of the property
train_station_id (15/100)	The closest train station to the property that has a direct trip to the Southern Cross Railway Station. A direct trip is a trip that, there are no connections (transfers) in trip from the origin to the destination. **Default value: 0**
distance_to_train_station (5/100)	The Haversine distance from the closest train station to the property that has a direct trip to the Southern Cross Railway Station. **Default value: 0**
travel_min_to_CBD (20/100)	The average travel time (minutes) from the closest train station (regional/metropolitan) that has a direct trip to the "Southern Cross Railway Station" on weekdays (i.e. Monday-Friday) **departing** between 7 to 9:30 am. For example, if there are 3 direct trips departing from the closest train station to the Southern Cross Railway Station on weekdays between 7-9:30 am and each takes 6, 7, and 8 minutes respectively, then the value of this column for the property should be (6+7+8)/3.). **Default value: 0**
over_priced? (10/100)	A boolean variable indicating whether or not the property price is higher than the median price of similar properties (with respect to bedrooms, bathrooms, parking_space, and property_type attributes) in the same suburb on the year of selling. **Default value: -1**
crime_A_average (7/100)	The average of type A crime for three years prior to selling **in the local government area** of the property as property. For example, if a property was sold in 2016, then you should calculate the average of the crime type A for 2013, 2014 and 2015. **Default value: -1**
crime_B_average (7/100)	The average of type B crime for three years prior to selling **in the local government area** as the property. For example, if a property was sold in 2016, then you should calculate the average

	of the crime type B crime for 2013, 2014 and 2015. **Default value: -1**
crime_C_average (6/100)	The average of type C crime for three years prior to selling **in the local government area** as the property. For example, if a property was sold in 2016, then you should calculate the average of the crime type C for 2013, 2014 and 2015. **Default value: -1**

Task 2: data reshaping (15%)

In this task, you need to study the effect of different normalization/transformation methods (i.e. standardization, min-max normalization, log, power, and root transformation) on *Rooms*, *crime_C_average, travel_min_to_CBD*, and *property_age* attributes. You need to observe and explain their effect assuming that we want to build a linear model on **price using these attributes** as the predictors of the linear model and recommend which one(s) you think would work better on this data. When building the linear model, the same normalization/transformation method can be applied to each of these attributes.

Task 3: Documentation and Methodology (25%)

The main focus of the documentation would be on the quality of your explanation of finishing these tasks. Your notebook file should be in a good format with proper sections and subsections.

Note 1: the output CSV file must have the exact same columns as specified on the schema. If you decide not to calculate any of the required attributes, then you must have a column for that attribute in your final data frame with the default value as the value of all the rows. Please note that the output file which is not in the correct format, as specified in the integrated schema, won't be marked.
Note 2: the radius of the earth is 6378 km.
Note 3: In table 2, numbers in front of some of the rows in the format of (a/b) are the allocated mark associated with that attribute. For example, the "suburb" attribute carries 20% of the total mark of task 1. Please note that 10% of the total marks for task 1 are marked on any other issue that may occur during the data integration process.
Note 4: You can only use the *vic_suburb_boundary.zip* file to extract the suburb name of the property. Using other external datasets or packages (e.g., Geopy) to get the suburb information directly will be penalised (this will result in 0 marks for the suburb attribute).
Note 5: for more info about GTFS data please visit here, here, and here.