FIT5196-S2-2019 assessment 4

This is a group assessment and worth 20% of your total mark for FIT5196.

Due date: **11:55 PM, Sunday, October 27, 2019**

For this assessment, you are required to write Python (Python 2/3) 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**	**Other Deliverables**
<GroupName>.csv vic_suburb_boundary.zip gtfs.zip Crimebylocation.xlsx Council.txt schools.xml	<GroupName>_solution.csv	<GroupName>_ass4.ipynb <GroupName>_dec-form.pdf <GroupName>_ref-diary.pdf

You are given multiple datasets in various formats and the task is about creating a housing information dataset in Victoria, Australia. Your assessment is to perform the following tasks.

Task 1: Data Integration (65%)

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 (15/100)	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 space of the property
LandSize	The area of the property
Age	The age of the property at the time of selling
Latitude	The Latitude of the property
Longitude	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 where there are no connections (transfers) in the trip from the origin to the destination. **Default value: -1**
distance_to_train_station (5/100)	The direct distance from the closest train station (identified above) to the property. **Default value: -1**
travel_min_to_CBD (15/100)	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: -1**
over_priced? (4/100)	A boolean feature indicating whether or not the price of the property 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 (3/100)	The average of type A crime **in the local government area** the property belongs to, in the three years prior to selling the property as the property. For example, if a property is 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 (3/100)	The average of type B crime **in the local government area** the property belongs to, in the three years prior to selling the property as the property. For example, if a property is sold in 2016, then you should calculate the average of the crime type B for 2013, 2014 and 2015. **Default value: -1**
crime_C_average (3/100)	The average of type C crime **in the local government area** the property belongs to, in the three years prior to selling the property as the property. For example, if a property is sold in 2016, then you should calculate the average of the crime type C for 2013, 2014 and 2015. **Default value: -1**
closest_primary_school (3/100)	The name of the closest primary school to the property. **Default value: "not available"**
distance_to_closest_primary (3/100)	The direct distance between the property and the closest primary school. **Default value: -1**
primary_school_ranking (11/100)	The ranking of the closest primary school to the property as scraped from http://www.schoolcatchment.com.au/?p=12301 If the school is not listed, the value of this field should be set to "not ranked". **Default value: -1**
closest_secondary_school (3/100)	The name of the closest secondary school to the property. **Default value: "not available"**
distance_to_closest secondary (3/100)	The direct distance between the property and the closest secondary school. **Default value: -1**
secondary_school_ranking (12/100)	The ranking of the closest secondary school to the property as scraped from https://sites.google.com/a/monash.edu/secondary-school-ranking/ If the school is not listed, the value of this field should be set to "not ranked". **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 predictors of the linear model and recommend which one(s) do 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 (20%)

The main focus on the documentation would be on the quality of your explanation on finishing these tasks. Your notebook file should be on a decent 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 output file which is not in a correct format, as specified in the integrated schema, won't be marked. Please be careful about naming your columns exactly as the requirement to avoid losing any marks.
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 15% of the total mark of task 1. Please note that 2% of the total marks for task 1 is marked on any other issue that you identify and fix 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 directly get the suburb information will be penalized (this will result in 0 marks for the suburb attribute).

Note 5: for more info about GTFS data please visit here , here, and here.

Note 6: You can use **difflib** python library for sequence matching and similarity checking.

Note 7: While retrieving school rankings, in case you find any duplicated ranks, always pick the better ranking for a school (note that rank 1 is better than rank 10, i.e. the lower the number, the better the rank)