FIT5196-S2-2020 assessment 3

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

Due date: 11:55 pm, November 18

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	Jupyter notebook and pdf
<student_no>.zip, Vic_suburb_boundary.zip, GTFS_Melbourne_Train_Information.zip	<student_no>_A3_solution.csv	<student_no>_ass3.ipynb <student_no>_ass3.pdf

The pdf file should be generated from your jupyter notebook file (after clearing all the cells output) and it will be used for plagiarism checks via turnitin.

Each of you is given 7 datasets in various formats and the 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 (60%)

In this task, you are required to integrate the input datasets (i.e., 7 datasets including **hospitals**, **supermarkets**, **shopping centers**, **real estate** files (one XML and one Json), **Vic_suburb_boundary**, and **GTFS_Melbourne_Train_Information** files) 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”
price	The property price
property_type	The type of the property
year	Year of sold
bedrooms	Number of bedrooms
bathrooms	Number of bathrooms
parking_space	The number of parking space of the property
Shopping_center_id (5%)	The closest shopping center to the property. **Default value: “not available”**
Distance_to_sc (5%)	The Euclidean distance from the closest shopping center to the property. **Default value: 0**
Train_station_id (10%)	The closest train station to the property. **Default value: 0**
Distance_to_train_station (5%)	The Euclidean distance from the closest train station to the property. **Default value: 0**
travel_min_to_CBD (15%)	The average travel time (minutes) from the closest train station to the “Flinders street” station on weekdays (i.e. Monday-Friday) **departing** between 7 to 9 am. For example, if there are 3 trip departing from the closest train station to the Flinders street 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 (6+7+8)/3. If there are any direct transfers between the closest station and Flinders street station, only the average of direct transfers should be calculated. **Default value: 0**
Transfer_flag (15%)	A Boolean attribute indicating whether there is a direct trip to the Flinders street station from the closest station between 7-9am on the weekdays. This flag is 0 if there is a direct trip (i.e. no transfer between trains is required to get from the closest train station to the Flinders station) and 1 otherwise. **Default value: -1**
Hospital_id (5%)	The closest hospital to the property. **Default value: “not available”**
Distance_to_hospital (5%)	The Euclidean distance from the closest hospital to the property. **Default value: 0**

Supermarket_id (5%)	The closest supermarket to the property. **Default value: "not available"**
Distance_to_supermaket (5%)	The Euclidean distance from the closest supermarket to the property. **Default value: 0**

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 **"price"**, **"Distance_to_sc"**, **"travel_min_to_CBD"**, and **"Distance_to_hospital"** attributes and observe and explain their effect assuming **we want to develop a linear model to predict the "price" using "Distance_to_sc", "travel_min_to_CBD", and "Distance_to_hospital" attributes**. The linear regression assumptions that you need to study in this task are: Normality and Linearity.

Task 3: Documentation (20%)

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.

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: if you decide not to calculate any of the required columns, then you must 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 3: 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.

Note 4: the radius of the earth is still 6378 km!

Note 5: In table 2, 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.