FIT5196-S1-2019 assessment 2

This is an individual assessment and worth 35% of your total mark for FIT5196.

Due date: **11:55 pm, 19 May 2019.**

Data Cleansing (%60)

For this assessment, you are required to write Python (Python 2/3) code to analyze your dataset, find and fix the problems in the data. The input and output of this task are shown below:

Table 1. The input and output of the task

Input	Output	Jupyter notebook
<student_no>_dirty_data.csv <student_no>_outliers.csv <student_no>_missing_value.csv	<student_no>_dirty_data_solution.csv <student_no>_outliers_solution.csv <student_no>.missing_value_solution.csv	<student_no>_ass2.ipynb

Exploring and understanding the data is one of the most important parts of the data wrangling process. You are required to perform both graphical and non-graphical EDA methods to understand the data first and then find the data problems. You are required to:

- Detect and fix errors in *<student_no>_dirty_data.csv*.
- Detect and remove outliers in *<student_no>_outliers.csv*.
- Impute the missing values in *<student_no>_missing_value.csv.*

As a starting point, here is all we know about the dataset in hand:
The dataset is about Uber Ridesharing data in **Victoria, Australia**. The description of each data column is shown in Table 2.

Table 2. Description of the columns

COLUMN	DESCRIPTION
Id	A unique id for the journey
Uber type	A categorical attribute for the type of the journey namely Uber pool, Uberx, Uber black. All we know is that the cost of these types of journeys may be different.
Origin region	A categorical attribute representing the region for the origin of the journey

Destination region	A categorical attribute representing the region for the destination of the journey
Origin latitude	Latitude of the origin coming from *nodes.csv* file
Origin longitude	Longitude of the origin coming from *nodes.csv* file
Destination latitude	Latitude of the destination coming from *nodes.csv* file
Destination longitude	Longitude of the destination coming from the *nodes.csv* file
Journey Distance	The shortest path, in meters, between the origin and the destination with respect to the nodes.txt and the edges.txt files. Dijkstra algorithm can be used to find the shortest path between two nodes in a graph. Reading materials can be found here.
Departure date	Date of the departure. We know that the price is different on weekdays compared to weekends.
Departure time	Time of the departure. We know that the Uber company has a specific rule to define a discrete number for morning (i.e. 0) (6:00:00 - 11:59:59), afternoon (i.e. 1) (12:00:00 - 20:59:59), and night (i.e. 2) (21:00 - 5:59:59) to calculate the fare.
Travel time	Travel time (i.e., duration) of the journey in seconds. Note that road types have their own speed limit in the *edges.csv* file and the car always travel with the exact speed limit.
Arrival time	The time of the arrival
Fare$	The fare of the journey. We know that the fare has a linear relation with some of the attributes of the dataset.

Notes:

1. The output *csv* files **must** have the exact same columns as the input.
2. There is at least one error in the dataset from each category of the data anomalies (i.e., syntactic, semantic, and coverage).
3. No rows carry more than one error.
4. There is no error other than outliers in the file *<student_no>_outliers.csv*. Similarly, there is no error other than missing value problems in the file *<student_no>_missing_value.csv.*
5. The radius of the Earth is 6378.0km.
6. Each Uber type has its own method of calculating the fare.
7. As EDA is part of this assessment, no further information will be given publicly regarding the data. However, you can brainstorm with the teaching team during tutorials and consultation sessions.

Methodology (%20)

The report should demonstrate the methodology (including all steps) to achieve the correct results.

Documentation (%20)

The cleaning task must be explained in a well-formatted report (with appropriate sections and subsections). Please remember that the report must explain the complete EDA to examine the data, your methodology to find the data anomalies and the suggested approach to fix those anomalies.