FIT5196-S2-2018 assessment 2

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

Due date: 11:55 pm, Wednesday, 3 October 2018

Data Cleansing (%70)

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>.csv	<student_no>_solution.csv	<student_no>_ass2.ipynb

Exploring and understanding the data is one of the most important parts in 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. However, as a starting point, here is all we know about the dataset in hand:
The dataset is about delivering packages using drones 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 delivery
Drone type	A categorical attribute for the type of the drone. We know that each type of drone has three phases of flight (namely *takeOff*, *onRoute*, and *Landing*). The drone may have different speeds at different phases. *takeOff* and *Landing* phases only take five minutes.
Post type	A categorical attribute for the type of delivery (0:normal, 1:express)
Package weight	The weight of the package

Origin region	A categorical attribute representing the region for the origin of the delivery
Destination region	A categorical attribute representing the region for the destination of the delivery
Origin latitude	Latitude of the origin
Origin longitude	Longitude of the origin
Destination latitude	Latitude of the destination
Destination longitude	Longitude of the destination
Distance	Distance of the journey
Departure date	Date of the departure
Departure time	Time of the departure. We know that the delivery company has a specific rule to define morning (6:00:00 - 11:59:59), afternoon (12:00:00 - 20:59:59), and night (21:00 - 5:59:59)
Travel time	Travel time (i.e., duration) of the journey
Delivery time	The time of the delivery
Delivery price	Delivery fare. We know that the fare has a linear relation with some of the attributes of the dataset.

Note 1: the output *csv* file must have the exact same columns as the input.
Note 2: the radius of the earth is 6378 km.
Note 3: 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.
Note 4: there is at least one error in the dataset from each category of the data anomalies (i.e., syntactic, semantic, and coverage).

Documentation (%30)

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.