FIT5196-S1-2022 assessment 2

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

Due date: Please check Moodle

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 files**	
<student_id>_dirty.csv <student_id>_outlier.csv <student_id>_missing.csv countries.csv	<student_id>_dirty_solution.csv <student_id>_outlier_solution.csv <student_id>_missing_solution.csv	<student_id>_ass2.ipynb <student_id>_ass2.pdf

Note1: The pdf file is the same as the notebook file but exported as "pdf" file (make sure to clear the output cells before exporting it)

Note2: All output files EXCEPT the pdf file must be zipped into a file named *<student_id>_ass2.zip*

Note3: <student_id> is to be replaced with your student_id

Note4: Each student can find their three input files and the supplementary file on the shared google drive.

Exploring and understanding the data is one of the most important parts of the data wrangling process. You are required to perform graphical and/or 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_id>_dirty.csv*
- Detect and **remove** outlier rows in *<student_id>_outlier.csv*
  (outliers are to be found w.r.t. *price* attribute)
- Impute the missing values in *<student_id>_missing.csv*

As a starting point, here is what we know about the dataset on hand:

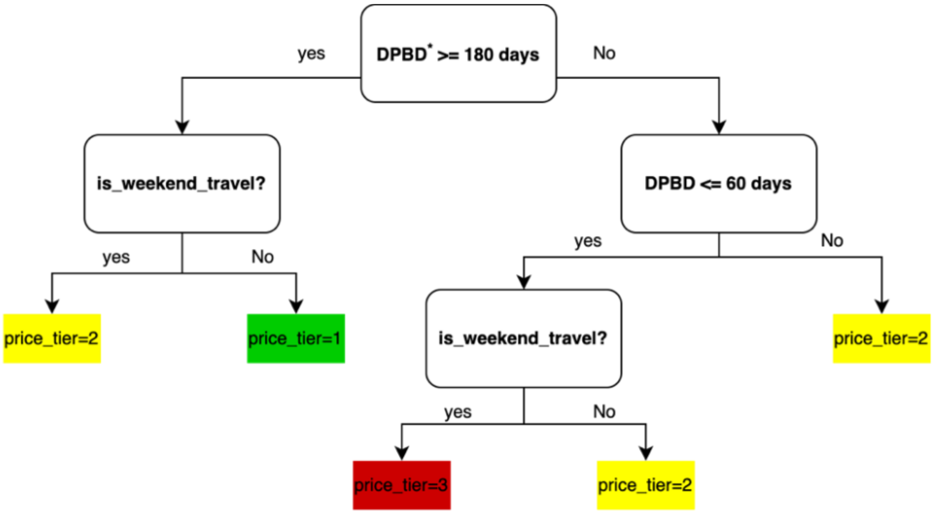
The dataset contains flight ticket bookings from three different airlines.
Each instance of the data represents a single ticket booking. The description of each data column is shown in Table 2.

Table 2. Description of the columns

COLUMN	DESCRIPTION
ticket_id	a unique identifier for each booking
airline_flight	a string denoting airline_name\airline_short_form–flight_number
name	a string denoting the title, firstname, and lastname of the traveller
gender	a string denoting gender [M, F] for male, female respectively.
dob	traveller date of birth (YYYY-MM-DD)
dop	date of ticket purchase (YYYY-MM-DD)
dod	date of departure (YYYY-MM-DD)
is_adult	a flag denoting if the traveller is 16 years or above on the date of departure
from	a string denoting the country of departure
to	a string denoting the destination country
distance	a float denoting the arc distance (in KM) between the two countries (radius of earth = 6378KM)
num_stops	an integer denoting the number of stops/transits during the flight
has_loyalty	a flag denoting whether the traveller has a loyalty program membership
price_tier	an integer in [1,2,3] denoting the price tier of the ticket (see notes)
price	a float denoting the purchase price of the ticket

Notes:

1. The output *csv* files **must** have the exact same columns as the respective input files.
2. In the file *<student_id>_dirty.csv,* any row can carry no more than one anomaly. (i.e. there can only be one issue in a single row.)
3. All anomalies have one and only one possible fix.
4. There are no data anomalies in the file *<student_id>_outlier.csv* except for outliers. Similarly, there are no data anomalies in *<student_id>_missing.csv* except for missing data.
5. The countries information is provided in the countries.csv file.
6. The price tier decides how the ticket is priced (see 7), to calculate the price tier, use the decision tree given in the figure below.



* DPBD : number of days from purchase date to departure date

7. The price of the ticket is calculated using a linear model (**different for each airline**) which depends on the following factors:
   - The interaction between price_tier and distance (i.e. price_tear * distance)
   - The number of stops/transits during the flight
   - Whether the traveller is an adult on the day of departure (i.e. 16 years or above).

   It is recommended to use **sklearn.linear_model.LinearRegression** for solving the linear model as demonstrated in our tutorials.
8. If the customer is a member of a loyalty program, the price of the ticket has a 10% discount applied.
9. ticket_id, name, from, to, num_stops, and price, in dirty data, are error-free.

As EDA is part of this assignment, no further information will be provided. You are encouraged to reach out to the teaching team for further discussions and/or clarifications.

Methodology (25%)

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

Documentation (15%)

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.