# DATA8001 Assignment 1

## Summary

The DATA8001 Assignment 1 is worth 50% of your overall module score.

Download the zip file from Canvas corresponding to your student id and unzip the contents into your local assignment folder and ensure your files are similar to Figure 1.

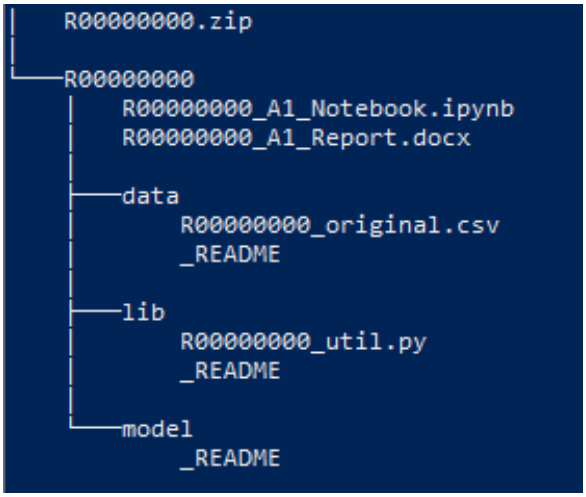


*Figure 1 - Example Assignment Folder & Files*

## Assignment Sections (50%)

### Data ETL – 10%

Clean the dataset provided: data/R00000000_original.csv and save as data/R00000000_processed.csv replacing R00000000 with your CIT student number.

All code required to reproduce the data ETL process should be placed in the Python library file (at the bottom where indicated): lib/R00000000_util.py and able to be called from the Jupyter Notebook: R00000000_A1_Notebook.ipynb.

#### Original Data Headings

| Column Name | Column Description |
|---|---|
| car_reg | the car registration plate |
| purchase_date | the purchase date of the car |
| county | the county car was purchased & registered |
| make | the car manufacturers name |
| model | the car model name |
| type | the type of car (e.g., saloon, hatchback etc.) |
| colour | the colour of the car |
| tax_band | the tax band of the car |
| price | the purchase price of the car in Euros |

DATA8001 Assignment 1

## Processed Data Headings & Expected Data Types

| Column Name | Column Description | Data Type |
|---|---|---|
| car_reg | Cleaned car registration plate | String (uppercase) |
| purchase_date | Cleaned purchase date of the car | Datetime |
| year | The year the car was purchased | Int |
| month | The month the car was purchased | Int |
| county | Cleaned county name | String (uppercase) |
| make | Cleaned car manufacturers name | String (uppercase) |
| model | Cleaned car model name | String (uppercase) |
| type | Cleaned car type | String (uppercase) |
| colour | Cleaned colour of the car | String (uppercase) |
| tax_band | Cleaned tax band of the car | String (uppercase) |
| price | the purchase price of the car in Euros | Float |

*Example*

| | car_reg | purchase_date | year | month | county | make | model | type | colour | tax_band | price |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 201-D-727 | 2020-02-10 | 2020 | 2 | DUBLIN | OPEL | CROSSLAND X | SUV | RED | A | 23323 |
| 1 | 201-D-3352 | 2020-01-17 | 2020 | 1 | DUBLIN | MAZDA | MAZDA6 | ESTATE | WHITE | B | 40627 |

## Data Visualisation – 10%

Load the processed dataset (data/R00000000_processed.csv) into the assignment notebook: R00000000_A1_Notebook.ipynb and answer the 5 questions including 1 (& only 1) visualisation of your choice that best answers each question. Show your workings in the Jupyter Notebook for each question.

## Data Modelling – 10%

Create a Linear Regression model and any transformations required to give your model the best accuracy. Using the Python class provided in lib/R00000000_util.py, save the object to the model folder as: model/R00000000.pkl.

All code required to reproduce the modelling process should be placed in the Python library file: lib/R00000000_util.py and able to be called from the Jupyter Notebook: R00000000_A1_Notebook.ipynb.

The pickled model file should be loaded and called from the Jupyter Notebook and available to process unseen test data including any transformations required to ensure the model works. **Note:** the unseen test data will have the same headings & datatypes as your data/R00000000_processed.csv file.

## Report & Questions (15%)

Write a max 2-page report outlining the steps taken to complete the assignment. Identify any areas you feel are worth mentioning during the ETL, visualisation of modelling steps including any insights developed.

Answer 2 exam type questions (max 300 words) each. Note – due to the “open-book” nature of this assignment, a clean, concise and well-thought-out answer of your “own” viewpoint is expected, this is not a “cut and paste” exercise!

## Presentation (5%)

Presentations for the assignment will take place on Tue 20th April 2021 between 6pm and 10pm. Each student has 5 minutes to present their work. How you demonstrate your work is entirely up to you (e.g., PowerPoint, Jupyter notebooks, videos, mimes etc.). Only students present on the evening can be scored!

**Note:** DO NOT submit any PowerPoint files as part of your project submission, they will not be graded.

# Submission Details

Assignments are due to be uploaded as a zip file via CIT Canvas no-later than **5pm on Monday 19th April 2021**.

Students should upload a zip file with the same name as the downloaded zip file (e.g., R00000000.zip) containing their completed work containing ONLY the folders & files listed in Figure 2.

## Files:

- **R00000000_A1_Notebook.ipynb** – completed notebook to call ETL process, visuals and answers and modelling.
- **R00000000_A1_Report.docx** – 2-page report and 2 answer exam type questions
- **data/R00000000_processed.csv** – clean dataset
- **lib/R00000000_util.py** – all the Python code required to recreate your work.
- **models/R00000000.pkl** – your pickled model object (ML model and transformations)

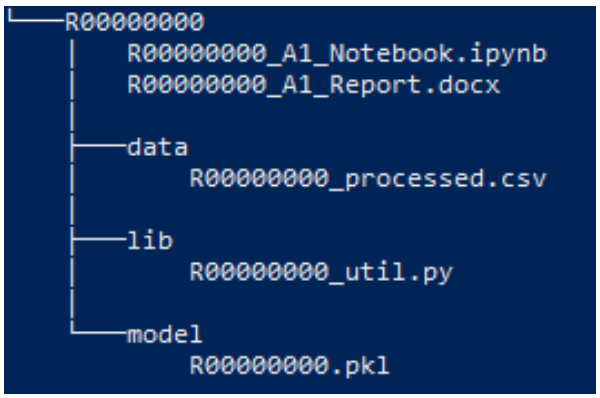


*Figure 2 - Example submission Folder & Files*