Intro

Data is the fundamental source of information in every area of life. It describes the reality surrounding us and allows us to know it better. Data can come from various sources, such as machine sensors, generators, software, databases, files and so on.

Data engineering refers to the techniques aimed at collecting, storing, managing and transforming raw data into a usable format for further analysis. This is typically a multi-step process. Before we can use the collected data, we must first import, analyze and preprocess it in the right way.

Preparing good quality data is usually time-consuming. However, it's worth investing the time to perform this. High quality data will allow for better analysis, visualization and overall understanding of the object being studied. This also enables the creation of higher quality object models, for example for machine learning purposes.

Theoretical Background

Data engineering is very closely related to the language of technical computing (LTC) workflow which is illustrated in figure 1. The presented workflow can be divided into 3 main stages: Access to data (import and collect data from various sources), Explore & Discover (analysis, processing and preparing for sharing), and Sharing results (documentation, data, code or deployment). MATLAB enables each step of the presented LTC workflow to be implemented using dedicated features, functions and interactive tools.

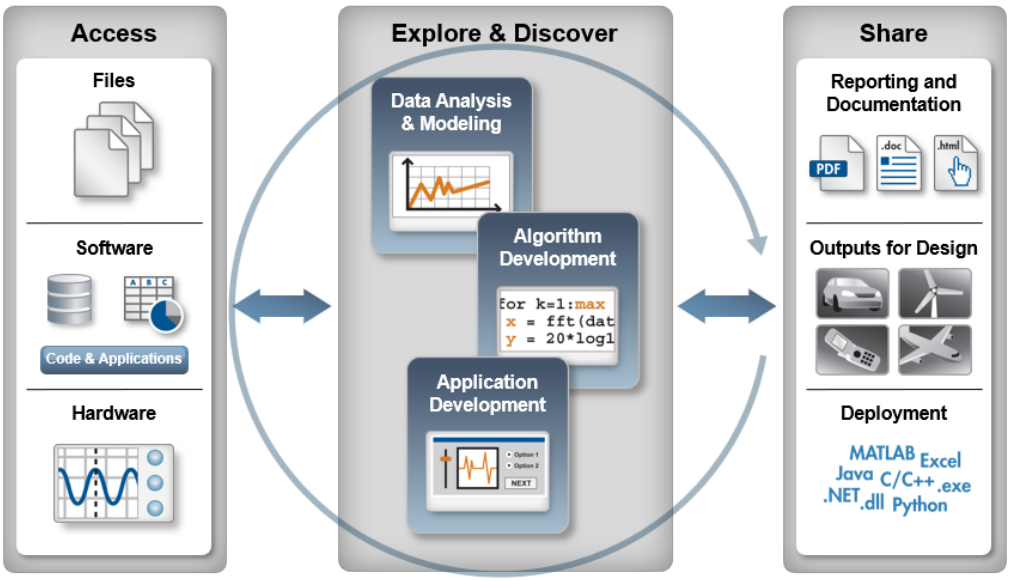

Fig. 1: Language of technical computing (LTC) workflow

Data can be stored in many formats. One of the most popular is CSV (comma-separated values) file format. In this type of file the data is organized in such a way that subsequent variables are stored in columns, while subsequent values (probes) are stored in rows and are usually separated by commas.

Data preprocessing and analyzing is one of the most important part of data engineering. In this part the following tasks are usually performed:

- data visualization and content analysis,
- improving data quality by filling in missing data, cleaning outliers and removing any noise or trends,
- detecting and assessing relationship between data and selecting the most valuable variables.

Data visualization, content analysis, variables relationship and improvement of their quality are based largely on statistical analyses, from basic computations such as min, max, average or standard deviation, to more advanced techniques such as correlation, interpolation, data prediction and modeling. When filling in missing data that depend on several other variables, regression models can be used. One of the common used models is the linear regression model. Filling in missing data can prevent an entire data record from being deleted due to a single missing value.

One of the most convenient ways to store and work with data is the tabular format. Each column corresponds to a specific variable, and the rows contain their values. A single table can contain various types of data, such as numeric, string, logical or categorical. Working with tables also allows you to easily manage data, export it and share with others.

Goal

Import the external data, process it according to the given requirements, save in tabular format and export to a shareable file.

Provided files

DataEngineering_input.csv – file containing data for import and processing.

Task description

The whole task is divided into two parts, i.e. Part I and Part II, described below. The data you will process according to the requirements given below must be finally saved as a **table**.

Follow the steps given below to complete the task.

Part I

1. Import data from an external comma-separated values (CSV) file named ***DataEngineering_input.csv*** into the MATLAB workspace as a **table**.
2. The following table variables (columns): *MPG*, *Acceleration* and *Weight* must not contain any outliers. When filling in outliers in these variables use one of the common MATLAB methods: *center, clip, previous, next, nearest, linear* or *spline*. When detecting outliers and verifying the filling in results, use the **default** detection method in MATLAB. Do not fill possible outliers in any other variables.
3. There are some missing data in the table. Firstly, fill in missing data in *Horsepower* variable. When filling in missing data use one of the common MATLAB methods:

previous, next, nearest, linear, spline, pchip or *makima*, but try to use the most appropriate method to the characteristics of the *Horsepower* values (e.g. linear, non-linear). To obtain the correct result reorganize the data in the *Horsepower* **in a simple way before filling it in**. Hint: *Horsepower* and *Displacement* variables are strongly positive correlated. For now, **do not** fill in possible missing data in the remaining variables. You will address with this matter later.

4. The *Region* variable must have 3 region types: *Europe*, *Japan* or *USA*.
5. The table should contain a new variable called *Displacement_ccm* with values in cubic centimeters calculated from the data in the *Displacement* variable. Use 1 inch = 2,54 cm for calculation.
6. The table should also contain a new variable named *MPG_Pred*. This variable should contain *true* entries in these rows where the *MPG* variable has *NaN* entries, and *false* entries in these rows where the *MPG* variable has numeric values.
7. The table should not contain the *Time_period* variable.
8. The table must contain the data processed with respect to the recommendations given above and has variables arranged in the following order and type:
   Region – type 'categorical'
   Origin – type 'cell'
   Manufacturer – type 'cell'
   Model_Year – type 'double'
   Horsepower – type 'double'
   Displacement – type 'double'
   Displacement_ccm – type 'double'
   MPG – type 'double'
   MPG_Pred – type 'logical'
   Acceleration – type 'double'
   Weight – type 'double'
   Cylinders – type 'double'
   The table size should be: 402 rows and 12 variables (columns).
9. The table cannot contain any missing data **except** *MPG* variable – this will be the aim of the Part II of the task.
10. The processed data must be saved in a **table** named: ***DataEng_output_table***

Part II

There are still some missing data in the *MPG* variable. Since this variable depends on several factors, its values can be predicted using a linear regression model.

1. Create a **linear regression model** fit to the input data using:
   - *Displacement*, *Weight* and *Model_Year* as the predictors,
   - *MPG* as the response data.

   When creating a linear regression model use the MATLAB **default model specification** input argument.

2. Using created linear regression model **predict** the missing values in the *MPG* using default input arguments, i.e. model and **new** predictors. Enter predicted *MPG* values in the table instead of missing values.

Finally, the team must save the processed data stored in the table ***DataEng_output_table*** to a file named ***DataEngineering_output.mat*** (MAT-file format).
A team can complete only Part I of the task and save the processed data to a table ***DataEng_output_table*** in the file ***DataEngineering_output.mat***, but then they may only receive a percentage of the points, as described in the TASK EVALUATION section below.
If a team completes both Part I and Part II of the task, they should save the results in the same table and file and then they can receive the maximum number of points.
The data to be processed is based on the MathWorks *carbig.mat* dataset containing measurements of cars from 1970–1982.

Task evaluation

Regarding the Part I of the task, the following will be assessed:

- whether the table is saved in the file with the required name and format,
- whether the processed data is saved in the table with the required name and size,
- correctness of the names and types of all variables, and their proper order in the table,
- cleaning outliers in the *MPG*, *Acceleration* and *Weight* variables,
- no missing data except MPG variable,
- proper fill in the missing data in the *Horsepower* with a MAPE error of no more than 1% compared to the reference data,

additionally, correctness of the data in the following variables:

- *Region*: 3 categories – *Europe*, *Japan* or *USA*,
- *Displacement_ccm*: a MAPE error of no more than 2% compared to the reference data,
- *MPG_Pred*: 4 *true* entries, other entries – *false*.

Regarding the Part II of the task, the following will be assessed:

- whether all predicted values in *MPG* do not exceed an absolute error of 10% compared to the reference values.

- If one or more of the requirements of Part I are not met, the team receives **0 points**.
- If all the requirements of Part I are met, the team receives **5 points**.
- If the Part I is done correctly, the team receives an additional **2 points** for correctly performing the Part II.
- In total the team can receives **7 points**.

Files to be uploaded

DataEngineering_output.mat – a MAT-file with *DataEng_output_table* table.

Useful documentation:

https://www.mathworks.com/help/matlab/ref/isoutlier.html
https://www.mathworks.com/help/matlab/ref/filloutliers.html
https://www.mathworks.com/help/matlab/ref/fillmissing.html
https://www.mathworks.com/help/stats/fitlm.html
https://www.mathworks.com/help/stats/linearmodel.predict.html