**DATA DESCRIPTION**

Houses.csv ~500Kb file with house sales data from yeah 2006 to 2010. Some columns were artificially added do provide grounds for exercise tasks, thus doesn’t contain any meaningful information.

The data consists of two headers, body, and one footer (special symbol “!” exists before header/footer).

Bottom line of footer contains expected row count. Counting column header row as well.

Field delimiter is “|”.

**DATA CLEANING TASK**

1.      Create 2 directories in your laptop:

-        RAW/in: receives data from external source.

- RAW/<YYYY>/<MM>/<DD>/: stores validated/prepared data (where YYYY, MM and DD are exact date).

-        BASE/<YYYY>/<MM>/<DD>/: stores data after cleaning and standardization (where YYYY, MM and DD are exact date).

2. Download the RAW data csv files from: shared driveto **“**RAW/in**”**

**N.B.** For this task basic python functionality and datetime, pathlib libraries should be enough.

3.      Create Python scripts for data validation/preparation

- read data from “RAW/in”

-        HEADER should be deleted.

-        File content should be validated (number of rows is equal to the value in FOOTER)

-        FOOTER should be deleted.

- save data to “RAW/<YYYY>/<MM>/<DD>”

**N.B.** For data cleaning task we will use Python libraries: Pandas, datetime is required, numpy optional.

5.      Create Python scripts for data cleaning

- read data from “RAW/<YYYY>/<MM>/<DD>”

- Drop the columns which have only NA values (DealCode)

- Remove duplicated rows (Id 1450 - 1460)

**-** Date should be standardized (including character Date to Date): date format should be transformed to YYYY-MM-DD (ActiveDate, EndDate, DateSold)

- Normalize data type inside one column (int instead of string, etc.)

-        Replace wrong decimal delimiter (“,”, etc) to standard style (“.”) delimiter. (SelPrice)

- Put all values in “String” columns into standardized case format (lower case)

- Add default values for missing data: for string – ‘’, for int – 0, float – 0.0, Date – 1970-01-01

- save data to “BASE/<YYYY>/<MM>/<DD>

6.      Scheduling/automatization of data cleaning scripts

- Create wrapper script in Python (which will trigger Validation/Preparing and Cleaning in sequential order).

- Create Scheduling task with Cron job (should be executed on daily basis)