ASSIGNMENT 1: CAR PRICES

Projects form an important part of the education of software engineers. They form an active method of teaching, as defined by Piaget, leading to a "training in self-discipline and voluntary effort", which is important to software engineering professionals. Two purposes served by these projects are: education in professional practice, and outcome-based assessment.

Data cleaning or data scrubbing is one of the most important steps previous to any data decision-making or modelling process. Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

Data cleaning is the process that removes data that does not belong to the dataset or it is not useful for modelling purposes. Data transformation is the process of converting data from one format or structure into another format. Transformation processes can also be referred to as data wrangling, or data munging, transforming and mapping data from one "raw" data form into another format. **Essentially, real-world data is messy data and for model building: garbage data in is garbage analysis out**.

This practical assignment belongs to Data Science Master at the UPC, any dataset for modelling purposes should include a first methodological step on **data preparation** about:

- Removing duplicate or irrelevant observations
- Fix structural errors (usually coding errors, trailing blanks in labels, lower/upper case consistency, etc.).
- Check data types. Dates should be coded as such and factors should have level names (if possible, levels have to be set and clarify the variable they belong to). This point is sometimes included under data transformation process. New derived variables are to be produced sometimes scaling and/or normalization (range/shape changes to numeric variables) or category regrouping for factors (nominal/ordinal).
- Filter unwanted outliers. Univariate and multivariate outliers have to be highlighted. Remove register/erase values and set NA for univariate outiers.
- Handle missing data: figure out why the data is missing. Data imputation is to be considered when the aim is modelling (imputation has to be validated).
- Data validation is mixed of 'common sense and sector knowledge': Does the data make sense? Does the data follow the appropriate rules for its field? Does it prove or disprove the working theory, or bring any insight to light? Can you find trends in the data to help you form a new theory? If not, is that because of a data quality issue?

Data Description

100,000 UK Used Car Data set

This data dictionary describes data (https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes) - A sample of 5000 trips has been randomly selected from Mercedes, BMW, Volkwagen and Audi manufacturers. So, firstly you have to combine used car from the 4 manufacturers into 1 dataframe.

The cars with engine size 0 are in fact electric cars, nevertheless Mercedes C class, and other given cars are not electric cars, so data imputation is required.

manufacturer	Factor: Audi, BMW, Mercedes or Volkswagen
model	Car model
year	registration year
price	price in £
transmission	type of gearbox
mileage	distance used
fuelType	engine fuel
tax	road tax
mpg	Consumption in miles per gallon
engineSize	size in litres

This course project (Assignment 1) is concerned with numeric model building for scraped data of used cars, which have been separated into files corresponding to each car manufacturer (only Mercedes, BMW, Volkswagen and Audi cars are to be considered): Y- Price (Numeric Target).

Aim is to predict how much you should sell your old car. It involves a numeric outcome.
A random sample containing 5000 registers combining Audi, VW, Merc and BMW registers has to be retained by each group. Data from:
https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes

1. Determine if the response variable (price) has an acceptably normal distribution. Address test to discard serial correlation.
2. Indicate by exploration of the data which are apparently the variables most associated with the response variable (use only the indicated variables).
3. Define a polytomic factor f.age for the covariate car age according to its quartiles and argue if the average price depends on the level of age. Statistically justify the answer.
4. Calculate and interpret the anova model that explains car price according to the age factor and the fuel type.
5. Do you think that the variability of the price depends on both factors? Does the relation between price and age factor depend on fuel type?
6. Calculate the linear regression model that explains the price from the age: interpret the regression line and assess its quality.
7. What is the percentage of the price variability that is explained by the age of the car?

8. Do you think it is necessary to introduce a quadratic term in the equation that relates the price to its age?
9. Are there any additional explanatory numeric variables needed to the car price? Study collinearity effects.
10. After controlling by numerical variables, indicate whether the additive effect of the available factors on the price are statistically significant.
11. Select the best model available so far. Interpret the equations that relate the explanatory variables to the answer (rate).
12. Study the model that relates the logarithm of the price to the numerical variables.
13. Once explanatory numerical variables are included in the model, are there any main effects from factors needed?
14. Graphically assess the best model obtained so far.
15. Assess the presence of outliers in the studentized residuals at a 99% confidence level. Indicate what those observations are.
16. Study the presence of *a priori* influential data observations, indicating their number according to the criteria studied in class.
17. Study the presence of *a posteriori* influential values, indicating the criteria studied in class and the actual atypical observations.
18. Given a 5-year old car, the rest of numerical variables on the mean and factors on the reference level, what would be the expected price with a 95% confidence interval?
19. Summarize what you have learned by working with this interesting real dataset.

Data Preparation outline:

Univariate Descriptive Analysis (to be included for each variable**):**

- Original numeric variables corresponding to qualitative concepts have to be converted to factors.
- Original numeric variables corresponding to real quantitative concepts are kept as numeric but additional factors should also be created as a discretization of each numeric variable.
- Exploratory Data Analysis for each variables (numeric summary and graphic support).

Data Quality Report:

Per variable, count:

- Number of missing values
- Number of errors (including inconsistencies)
- Number of outliers
- Rank variables according the sum of missing values (and errors).

Per individuals, count:

- number of missing values
- number of errors,
- number of outliers
- Identify individuals considered as multivariant outliers.

Create variable adding the total number missing values, outliers and errors. Describe these variables, to which other variables exist higher associations.

- Compute the correlation with all other variables. Rank these variables according the correlation
- Compute for every group of individuals (group of age, size of town, singles, married, …) the mean of missing/outliers/errors values. Rank the groups according the computed mean.

Imputation:

- Numeric Variables
- Factors

Profiling:

- Target (age)