ASSIGNMENT 1: CANCER MORTALITY

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 modeling 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 modeling 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 means garbage out**.

This practical assignment belongs to Data Science Master at the UPC, any dataset for modeling 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. Data 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 in the 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 outliers.
- Handle missing data: figure out why the data is missing. Data imputation is to be considered when the aim is modeling (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?

Dataset Context and Contents

The Cancer Mortality dataset is for use in data science education. It can be found on the data.world website (https://data.world/exercises/linear-regression-exercise-1), it's been split to train and test samples. There are 1831 observations in the train dataset and 1216 in the test dataset. **The target variable is target_deathrate** .

Student team consists of 2/3 students. Contribution of each team member has to be included in the report.

- Hint: You have to retain all available numeric variables except those that are not useful in explaining the target.

Variables

avganncount	Mean number of reported cases of cancer diagnosed annually (2010-2015
avgdeathsperyear	Mean number of reported mortalities due to cancer
target_deathrate	**Response variable. Mean per capita (100,000) cancer mortalities**
incidencerate	Mean per capita (100,000) cancer diagnoses
medincome	Median income per county
popest2015	Population of county
povertypercent	Percent of population in poverty
studypercap	Per capita number of cancer-related clinical trials per county
binnedi	Median income per capita binned by decile
medianage	Median age of county residents
medianagemale	Median age of male county residents
medianagefemale	Median age of female county residents
geography	County name
percentmarried	Percent of county residents who are married
pctnohs18_24	Percent of county residents ages 18-24 highest education attained: less than high school
pcths18_24	Percent of county residents ages 18-24 highest education attained: high school diploma
pctsomecol18_24	Percent of county residents ages 18-24 highest education attained: some college
pcths25_over	Percent of county residents ages 25 and over highest education attained: high school diploma
pctbachdeg25_over	Percent of county residents ages 25 and over highest education attained: bachelor's degree
pctemployed16_over	Percent of county residents ages 16 and over employed
pctunemployed16_over	Percent of county residents ages 16 and over unemployed
pctprivatecoverage	Percent of county residents with private health coverage
pctprivatecoveragealone	Percent of county residents with private health coverage alone (no public assistance)

pctempprivcoverage	Percent of county residents with employee-provided private health coverage
pctpubliccoverage	Percent of county residents with government-provided health coverage
pctpubliccoveragealone	Percent of county residents with government-provided health coverage alone
pctwhite	Percent of county residents who identify as White
pctblack	Percent of county residents who identify as Black
pctasian	Percent of county residents who identify as Asian
pctotherrace	Percent of county residents who identify in a category which is not White, Black, or Asian
pctmarriedhouseholds	Percent of married households
birthrate	Number of live births relative to number of women in county

- Exploratory Data Analysis and Model Fitting should take **train sample** only.
- Create factors for retained qualitative variables. **Train and Test samples**.
- Determine if the response variable (deathrate) has an acceptably normal distribution.
- Address tests to discard serial correlation.
- Detect univariant and multivariant outliers and **retain** all of them in exploratory analysis.
- **Errors and missing** values (if any) detection. Apply an imputation technique for both **train and test** datasets, if needed.
- Preliminary exploratory analysis to describe observed relations has to be undertaken.
- If you can improve linear relations or limit the effect of influential data, you must consider suitable transformations for variables.
- Apart from the retained factor variables, you can consider other categorical variables that can be defined from categorized numeric variables. **Do not forget to implement new variable definitions in the test sample.**
- You must take into account possible **interactions** between categorical and numerical variables.
- When building the model, you should study the presence of **multicollinearity** and try to reduce their impact on the model for easier interpretation.
- You should build the model using a technique for selecting variables (removing no significant predictors and/or stepwise selection of the best models).
- The validation of the model has to be done with graphs and / or suitable tests to verify model assumptions.
- You must include the study of unusual and / or influential data.
- The resulting model should be interpreted in terms of the relationships of selected predictors and its effect on the response variable.
- You have to apply your final model to the **test sample** and roughly assess forecasting capability.