ASSIGNMENT 1: GEORGIA HOUSE PRICE

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 ***Industria 4.0, Modelización, Simulación y Materialización*** at the Talent School 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 Georgia House dataset is for use in data science education. It can be found on the Kaggle website (https://www.kaggle.com/datasets/yellowj4acket/real-estate-georgia). There are 6,168 observations in total that should be split into train and test dataset, but skipped in order to maintain the same sample for all the students. **The target variable is price.**

Student team size is recommended to be 3 students. Contribution of each team member has to be included in the report. Splitting into work and test data does not have to be considered in order to keep the set of observations for all students (Assignment Exam). Nevertheless, splitting process is recommended in any professional project.

Hints:

- You have to retain observation that were posted on **2021 only** to avoid any bias.
- Pay attention to the inconsistency of columns, especially those that have both numeric and categorical variables present in the dataset.
- Pay attention to number of unique values for each variable, some may not be useful in the analysis at all.
- Zero value for bathrooms is not acceptable!
- You must retain all available numeric variables **except** those that are not useful in explaining the target. livingArea and pricePerSquareFoot perfectly describe the target, you may use one in the model.
- Unreasonably huge values for some columns may considered as errors in the data and need proper consideration.
- Use **unique** observations only in your analysis.

Variables

```
Load dataset using:
df<- read.csv("RealEstate_Georgia.csv",header = T)

The dataset contains 6168 housing ads from different cities in different counties
                     of the Georgia state, there are posts from 2014 to 2021 but
                     majority of data are for 2021. For a fair analysis, the houses
                     that are posted in 2021 are taken into account.


df$datePostedString <- as.numeric(format(df$datePostedString, "%Y"))
df <- df[which(df$datePostedString ==2021),]
```

X	
id	
stateId	16 == georgia

countyId	county id
cityId	city id
country	USA only
datePostedString	published on:
is_bankOwned	offered by a bank?
is_forAuction	auctioned property?
event	different sales types
time	time collected
price	USD
pricePerSquareFoot	USD / sqft
city	city
state	state
yearBuilt	Year built
streetAddress	address
zipcode	Zipcode
longitude	longitude
latitude	latitude
hasBadGeocode	bad geocode?
description	free text description
currency	currency
livingArea	living are sqft
livingAreaValue	area in sqft
bathrooms	Number of bathrooms
bedrooms	Number of bedrooms
buildingArea	area
parking	has parking/binary
garageSpaces	Number of garage spaces
hasGarage	has garage /binary
levels	Levels of the building
pool	Has pool
spa	Has spa
isNewConstruction	Is it a new building?
hasPetsAllowed	Is pet allowed?
homeType	Home type
county	county

- Determine if the response variable (price) has an acceptably normal distribution.
- Address tests to discard serial correlation.
- Detect univariant and multivariant outliers and **remove them if it's needed** in explanatory analysis.
- **Errors and missing** values (if any) should be treated. Apply an imputation technique, 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. Variables with univariate outliers are candidates to be categorized.

- You may redefine categories of column **levels** to be used in the model.
- You must consider 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.