ASSIGNMENT 1: HOUSE 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 means garbage 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?

Dataset Context and Contents

The Ames Housing dataset was compiled by Dean De Cock for use in data science education. It can be found in the Kaggle website (https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview), there are 1460 observations in the train dataset and 1459 in the test dataset. **Target variable is SalePrice** .

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. You are allowed to select a subset of about 10 available factors.

Variables

```
MSSubClass: Identifies the type of dwelling involved in the sale.

        20   1-STORY 1946 & NEWER ALL STYLES
        30   1-STORY 1945 & OLDER
        40   1-STORY W/FINISHED ATTIC ALL AGES
        45   1-1/2 STORY - UNFINISHED ALL AGES
        50   1-1/2 STORY FINISHED ALL AGES
        60   2-STORY 1946 & NEWER
        70   2-STORY 1945 & OLDER
        75   2-1/2 STORY ALL AGES
        80   SPLIT OR MULTI-LEVEL
        85   SPLIT FOYER
        90   DUPLEX - ALL STYLES AND AGES
       120   1-STORY PUD (Planned Unit Development) - 1946 & NEWER
       150   1-1/2 STORY PUD - ALL AGES
       160   2-STORY PUD - 1946 & NEWER
       180   PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
       190   2 FAMILY CONVERSION - ALL STYLES AND AGES

MSZoning: Identifies the general zoning classification of the sale.

       A     Agriculture
       C     Commercial
       FV    Floating Village Residential
       I     Industrial
       RH    Residential High Density
       RL    Residential Low Density
       RP    Residential Low Density Park
       RM    Residential Medium Density

LotFrontage: Linear feet of street connected to property (numeric)

LotArea: Lot size in square feet (numeric)

Street: Type of road access to property

       Grvl  Gravel
       Pave  Paved

Alley: Type of alley access to property

       Grvl  Gravel
       Pave  Paved
       NA    No alley access
```

```
LotShape: General shape of property

       Reg   Regular
       IR1   Slightly irregular
       IR2   Moderately Irregular
       IR3   Irregular

LandContour: Flatness of the property

       Lvl   Near Flat/Level
       Bnk   Banked - Quick and significant rise from street grade to building
       HLS   Hillside - Significant slope from side to side
       Low   Depression

Utilities: Type of utilities available

       AllPubAll public Utilities (E,G,W,& S)
       NoSewrElectricity, Gas, and Water (Septic Tank)
       NoSeWaElectricity and Gas Only
       ELO   Electricity only

LotConfig: Lot configuration

       Inside       Inside lot
       Corner       Corner lot
       CulDSac      Cul-de-sac
       FR2          Frontage on 2 sides of property
       FR3          Frontage on 3 sides of property

LandSlope: Slope of property

       Gtl   Gentle slope
       Mod   Moderate Slope
       Sev   Severe Slope

Neighborhood: Physical locations within Ames city limits

       Blmngtn      Bloomington Heights
       Blueste      Bluestem
       BrDale       Briardale
       BrkSide      Brookside
       ClearCr      Clear Creek
       CollgCr      College Creek
       Crawfor      Crawford
       Edwards      Edwards
       Gilbert      Gilbert
       IDOTRR       Iowa DOT and Rail Road
       MeadowV      Meadow Village
       Mitchel      Mitchell

Names  North Ames
       NoRidge      Northridge
       NPkVill      Northpark Villa
       NridgHt      Northridge Heights
       NWAmesNorthwest Ames
       OldTown      Old Town
       SWISU South & West of Iowa State University
       SawyerSawyer
       SawyerW      Sawyer West
       Somerst      Somerset
       StoneBr      Stone Brook
       Timber       Timberland
       Veenker      Veenker

Condition1: Proximity to various conditions

       ArteryAdjacent to arterial street
       Feedr Adjacent to feeder street
```

```
       Norm  Normal
       RRNn  Within 200' of North-South Railroad
       RRAn  Adjacent to North-South Railroad
       PosN  Near positive off-site feature--park, greenbelt, etc.
       PosA  Adjacent to postive off-site feature
       RRNe  Within 200' of East-West Railroad
       RRAe  Adjacent to East-West Railroad

Condition2: Proximity to various conditions (if more than one is present)

       ArteryAdjacent to arterial street
       Feedr Adjacent to feeder street
       Norm  Normal
       RRNn  Within 200' of North-South Railroad
       RRAn  Adjacent to North-South Railroad
       PosN  Near positive off-site feature--park, greenbelt, etc.
       PosA  Adjacent to postive off-site feature
       RRNe  Within 200' of East-West Railroad
       RRAe  Adjacent to East-West Railroad

BldgType: Type of dwelling

       1Fam  Single-family Detached
       2FmConTwo-family Conversion; originally built as one-family dwelling
       Duplx Duplex
       TwnhsETownhouse End Unit
       TwnhsITownhouse Inside Unit

HouseStyle: Style of dwelling

       1StoryOne story
       1.5FinOne and one-half story: 2nd level finished
       1.5UnfOne and one-half story: 2nd level unfinished
       2StoryTwo story
       2.5FinTwo and one-half story: 2nd level finished
       2.5UnfTwo and one-half story: 2nd level unfinished
       SFoyerSplit Foyer
       SLvl  Split Level

OverallQual: Rates the overall material and finish of the house

       10    Very Excellent
       9     Excellent
       8     Very Good
       7     Good
       6     Above Average
       5     Average
       4     Below Average
       3     Fair
       2     Poor
       1     Very Poor

OverallCond: Rates the overall condition of the house

       10    Very Excellent
       9     Excellent
       8     Very Good
       7     Good
       6     Above Average
       5     Average
       4     Below Average
       3     Fair
       2     Poor
       1     Very Poor

YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)
```

RoofStyle: Type of roof

Flat	Flat
Gable	Gable
Gambrel	Gabrel (Barn)
Hip	Hip
Mansard	Mansard
Shed	Shed

RoofMatl: Roof material

ClyTile	Clay or Tile
CompShg	Standard (Composite) Shingle
Membran	Membrane
Metal	Metal
Roll	Roll
Tar&Grv	Gravel & Tar
WdShake	Wood Shakes
WdShngl	Wood Shingles

Exterior1st: Exterior covering on house

AsbShng	Asbestos Shingles
AsphShn	Asphalt Shingles
BrkComm	Brick Common
BrkFace	Brick Face
CBlock	Cinder Block
CemntBd	Cement Board
HdBoard	Hard Board
ImStucc	Imitation Stucco
MetalSd	Metal Siding
Other	Other
Plywood	Plywood
PreCast	PreCast
Stone	Stone
Stucco	Stucco
VinylSd	Vinyl Siding
Wd Sdng	Wood Siding
WdShing	Wood Shingles

Exterior2nd: Exterior covering on house (if more than one material)

AsbShng	Asbestos Shingles
AsphShn	Asphalt Shingles
BrkComm	Brick Common
BrkFace	Brick Face
CBlock	Cinder Block
CemntBd	Cement Board
HdBoard	Hard Board
ImStucc	Imitation Stucco
MetalSd	Metal Siding
Other	Other
Plywood	Plywood
PreCast	PreCast
Stone	Stone
Stucco	Stucco
VinylSd	Vinyl Siding
Wd Sdng	Wood Siding
WdShing	Wood Shingles

MasVnrType: Masonry veneer type

BrkCmn	Brick Common
BrkFace	Brick Face
CBlock	Cinder Block
None	None
Stone	Stone

MasVnrArea: Masonry veneer area in square feet

ExterQual: Evaluates the quality of the material on the exterior

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
Po	Poor

ExterCond: Evaluates the present condition of the material on the exterior

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
Po	Poor

Foundation: Type of foundation

BrkTil	Brick & Tile
CBlock	Cinder Block
PConc	Poured Contrete
Slab	Slab
Stone	Stone
Wood	Wood

BsmtQual: Evaluates the height of the basement

Ex	Excellent (100+ inches)
Gd	Good (90-99 inches)
TA	Typical (80-89 inches)
Fa	Fair (70-79 inches)
Po	Poor (<70 inches
NA	No Basement

BsmtCond: Evaluates the general condition of the basement

Ex	Excellent
Gd	Good
TA	Typical - slight dampness allowed
Fa	Fair - dampness or some cracking or settling
Po	Poor - Severe cracking, settling, or wetness
NA	No Basement

BsmtExposure: Refers to walkout or garden level walls

Gd	Good Exposure
Av	Average Exposure (split levels or foyers typically score average or above)
Mn	Mimimum Exposure
No	No Exposure
NA	No Basement

BsmtFinType1: Rating of basement finished area

GLQ	Good Living Quarters
ALQ	Average Living Quarters
BLQ	Below Average Living Quarters
Rec	Average Rec Room
LwQ	Low Quality
Unf	Unfinshed
NA	No Basement

BsmtFinSF1: Type 1 finished square feet (numeric)

BsmtFinType2: Rating of basement finished area (if multiple types)

```
       GLQ   Good Living Quarters
       ALQ   Average Living Quarters
       BLQ   Below Average Living Quarters
       Rec   Average Rec Room
       LwQ   Low Quality
       Unf   Unfinshed
       NA    No Basement

BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area (numeric)

TotalBsmtSF: Total square feet of basement area (numeric)

Heating: Type of heating

       Floor Floor Furnace
       GasA  Gas forced warm air furnace
       GasW  Gas hot water or steam heat
       Grav  Gravity furnace
       OthW  Hot water or steam heat other than gas
       Wall  Wall furnace

HeatingQC: Heating quality and condition

       Ex    Excellent
       Gd    Good
       TA    Average/Typical
       Fa    Fair
       Po    Poor

CentralAir: Central air conditioning

       N     No
       Y     Yes

Electrical: Electrical system

       SBrkr Standard Circuit Breakers & Romex
       FuseA Fuse Box over 60 AMP and all Romex wiring (Average)
       FuseF 60 AMP Fuse Box and mostly Romex wiring (Fair)
       FuseP 60 AMP Fuse Box and mostly knob & tube wiring (poor)
       Mix   Mixed

1stFlrSF: First Floor square feet (numeric)

2ndFlrSF: Second floor square feet (numeric)

LowQualFinSF: Low quality finished square feet (all floors)

GrLivArea: Above grade (ground) living area square feet (numeric)

BsmtFullBath: Basement full bathrooms

BsmtHalfBath: Basement half bathrooms

FullBath: Full bathrooms above grade (numeric)

HalfBath: Half baths above grade (numeric)

Bedroom: Bedrooms above grade (does NOT include basement bedrooms)

Kitchen: Kitchens above grade (numeric)

KitchenQual: Kitchen quality

       Ex    Excellent
```

```
       Gd    Good
       TA    Typical/Average
       Fa    Fair
       Po    Poor

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms) (numeric)

Functional: Home functionality (Assume typical unless deductions are warranted)

       Typ   Typical Functionality
       Min1  Minor Deductions 1
       Min2  Minor Deductions 2
       Mod   Moderate Deductions
       Maj1  Major Deductions 1
       Maj2  Major Deductions 2
       Sev   Severely Damaged
       Sal   Salvage only

Fireplaces: Number of fireplaces (numeric)

FireplaceQu: Fireplace quality

       Ex    Excellent - Exceptional Masonry Fireplace
       Gd    Good - Masonry Fireplace in main level
       TA    Average - Prefabricated Fireplace in main living area or Masonry
             Fireplace in basement
       Fa    Fair - Prefabricated Fireplace in basement
       Po    Poor - Ben Franklin Stove
       NA    No Fireplace

GarageType: Garage location

       2Types       More than one type of garage
       Attchd       Attached to home
       Basment      Basement Garage
       BuiltIn      Built-In (Garage part of house - typically has room above garage)
       CarPort      Car Port
       Detchd       Detached from home
       NA           No Garage

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

       Fin   Finished
       RFn   Rough Finished
       Unf   Unfinished
       NA    No Garage

GarageCars: Size of garage in car capacity (numeric)

GarageArea: Size of garage in square feet (numeric)

GarageQual: Garage quality

       Ex    Excellent
       Gd    Good
       TA    Typical/Average
       Fa    Fair
       Po    Poor
       NA    No Garage

GarageCond: Garage condition

       Ex    Excellent
       Gd    Good
       TA    Typical/Average
       Fa    Fair
```

```
       Po    Poor
       NA    No Garage

PavedDrive: Paved driveway

       Y     Paved
       P     Partial Pavement
       N     Dirt/Gravel

WoodDeckSF: Wood deck area in square feet (numeric)

OpenPorchSF: Open porch area in square feet (numeric)

EnclosedPorch: Enclosed porch area in square feet (numeric)

3SsnPorch: Three season porch area in square feet (numeric)

ScreenPorch: Screen porch area in square feet (numeric)

PoolArea: Pool area in square feet (numeric)

PoolQC: Pool quality

       Ex    Excellent
       Gd    Good
       TA    Average/Typical
       Fa    Fair
       NA    No Pool

Fence: Fence quality

       GdPrv Good Privacy
       MnPrv Minimum Privacy
       GdWo  Good Wood
       MnWw  Minimum Wood/Wire
       NA    No Fence

MiscFeature: Miscellaneous feature not covered in other categories

       Elev  Elevator
       Gar2  2nd Garage (if not described in garage section)
       Othr  Other
       Shed  Shed (over 100 SF)
       TenC  Tennis Court
       NA    None

MiscVal: $Value of miscellaneous feature

MoSold: Month Sold (MM)

YrSold: Year Sold (YYYY)

SaleType: Type of sale

       WD    Warranty Deed - Conventional
       CWD   Warranty Deed - Cash
       VWD   Warranty Deed - VA Loan
       New   Home just constructed and sold
       COD   Court Officer Deed/Estate
       Con   Contract 15% Down payment regular terms
       ConLw Contract Low Down payment and low interest
       ConLI Contract Low Interest
       ConLD Contract Low Down
       Oth   Other

SaleCondition: Condition of sale

       Normal       Normal Sale
```

Abnorml	Abnormal Sale - trade, foreclosure, short sale
AdjLand	Adjoining Land Purchase
Alloca	Allocation - two linked properties with separate deeds, typically condo with a garage unit
Family	Sale between family members
Partial	Home was not completed when last assessed (associated with New Homes)

SalePrice: Target variable (numeric)

The file contains some numeric variables (retain all of them) and many factors (restrict to 10 to reduce the effort).

- 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 (charges) has an acceptably normal distribution.
- Address tests to discard serial correlation.
- Detect univariant and multivariant outliers, errors and missing values (if any) and 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. 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.