**House Price Prediction**

The goal of this project is to estimate the house price of different available houses.

Each group will receive a dataset, which includes the same features. A detailed description of these features can be found in feature\_description.txt (Feature Description).

The dataset consists of synthetic data, so it was generated by us. This allowed us to incorporate many properties different real world datasets have. Hence, please keep in mind that not each of the listed houses might be constructed like described in the real world (e.g. 20 sqm with 3 bathrooms), but still look out for other outliers in the data (like mentioned in the first lecture) and otherwise general logic still applies to this dataset (e.g. a house that was built thousands of years ago, but has the same properties as a newer one, may not make sense)!

Your tasks include

* Exploratory Data Analysis and Preprocessing of the data
* Predicting the house prices
* Evaluating your model
* Answering the following Questions
  1. If I have a budget of €100,000, what kind of houses will I be able to buy?
  2. How much money can I save if I decide not to get a house with a pool?

You have to evaluate your model yourself, so consider what could be good and appropriate evaluation methods.

At the end of this semester, you have to give a short presentation on your work.

Your presentation should have exactly

* 1 or 2 EDA slides
* 1 Method slide
* 1 or 2 Evaluation slides
* 1 Discussion slide

Of course you're allowed to add a title slide if you want to. The presentation should take a maximum of 10 minutes.

Please hand in your presentation (Group\_xx.pdf) and your source code (Group\_xx.zip) until **23.06.2024**.

Have fun!

**Overview/Feature Description:**

Wie soll generell mit fehlenden Werten umgegangen werden?

* Entfernen
* Mittelwert
* Anpassen (Vorzeichen, Kommastellen, etc.)
* Achtung! Wenn bei Pool Nan, dann existiert einfach kein Pool! Diese Fälle dürfen nicht einfach entfernt werden!
* Bevor Reihen entfernt werden, müssen die Spalten entfernt werden!

Bedrooms - number of bedrooms

Bathrooms - number of bathrooms

* Wir könnten eine neue Spalte mit gesamter Zimmeranzahl erstellen (Bedrooms + Bathrooms)

SquareFootageHouse - the square footage of the house in square meters

* Es kommen negative Werte vor! Entweder das Vorzeichen ändern oder entfernen

Location - whether the house is urban, rural or suburban (Man könnte das Set anhand der Location analysieren)

* Achtung es gibt 4 Katgeorien durch einen Schreibfehler!
* ['Rural', 'Suburban', 'Suburbann', 'Urban']
* Suburban und Suburbann werden zusammengefasst

Age - the age of the house, all houses were built in the last 100 years

* Es kommen 12 negative Werte vor! Entweder das Vorzeichen ändern oder entfernen

PoolQuality - the quality of the pool, if it is 'None' then there is no pool

* ['Excellent', 'Good', 'Poor']

HasPhotovoltaics - if the house has photovoltaics

* Bool: Y/N

HeatingType - the heating type of the house

* Achtung es gibt 5 Kategorien!
* ['Electric', 'Electricity', 'Gas', 'Oil', 'Oil Heating']
* Werden zusammengefasst
* ['Electric', 'Gas', 'Oil']

HasFiberglass - if the house has the possibility of a fiberglass connection

* Bool: Y/N

IsFurnished - if the home is furnished

* Bool: Y/N

DateSinceForSale - the date the house was put on the market

* Vergleich ob Age < DateSinceForSale

HouseColor - the color of the house

* ['Gray', 'Green', 'White', 'Yellow']

PreviousOwnerName - the first name of the previous owner

* Finde ich unwichtig. Kann gelöscht werden

HasFireplace - presence of a fireplace

* Bool: Y/N

KitchensQuality - quality of the kitchen(s)

* ['Excellent', 'Good', 'Poor']

BathroomsQuality - quality of bathroom(s)

* ['Excellent', 'Good', 'Poor']

BedroomsQuality - quality of the bedroom(s)

* ['Excellent', 'Good', 'Poor']

LivingRoomsQuality - quality of the living room(s)

* ['Excellent', 'Good', 'Poor']

SquareFootageGarden - the square footage of the garden in square meters

* Wir könnten eine neue Variable “Gesamt Fläche” (Haus + Garten) erstellen

PreviousOwnerRating - the rating of the house by the previous owner

* From 1.0 to 4.97

HeatingCosts - the heating costs in Euro per month for the whole house

* Es kommen 5 negative Werte vor! Entweder das Vorzeichen ändern oder entfernen

Window\_model\_names - the name of the window model

* Insgesamt 995 verschiedene Kategorien => Keine Duplikate
* Bin dafür, dass diese Spalte entfernt werden sollte

Price - the price of the house in thousands of euros

**Tasks:**

* Exploratory Data Analysis and Preprocessing of the data
  + Augenschein Analyse und Lokalisierung von Ausreißern
  + Wie soll mit Ausreißern bzw. Unmöglichen Werten umgegangen werden? (Siehe oben bei der Overview)
    - Wie soll generell mit fehlenden Werten umgegangen werden?
      * Entfernen
      * Mittelwert
      * Anpassen (Vorzeichen, Kommastellen, etc.)
  + Können manche Spalten komplett entfernt werden? (z.B.: Fenster)
  + Bestehen Duplikate?
* Predicting the house prices
  + MLR
* Evaluating your model
* Answering the following Questions
  + If I have a budget of €100,000, what kind of houses will I be able to buy?
  + Spezifizierung der Hausart:
    - Möglichst groß (Fläche, Zimmer, etc.), neu, ...
    - Hat Zusätze: Pool, Photovoltaik, etc.
    - Je nach Ort (Urban, Subrban, ...)
  + How much money can I save if I decide not to get a house with a pool?
    - Vergleich von vergleichbaren Immobilien (Größe, Alter etc.) in Abhängigkeit von Pool/kein Pool

**Exploratory Data analysis:**

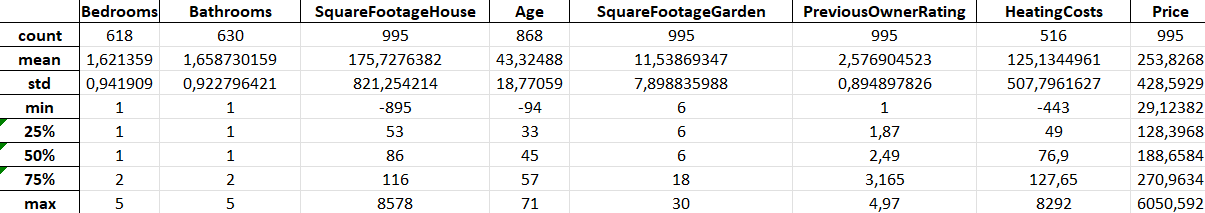
**Deskriptive Statistik:**

1000 Zeilen

23 Spalten

8 numerische Spalten

15 object Spalten



**Null Werte:**

Bedrooms 382

Bathrooms 370

SquareFootageHouse 5

Location 199

Age 132

PoolQuality 648

HasPhotovoltaics 84

HeatingType 5

HasFiberglass 5

IsFurnished 5

DateSinceForSale 5

HouseColor 13

PreviousOwnerName 151

HasFireplace 5

KitchensQuality 5

BathroomsQuality 5

BedroomsQuality 5

LivingRoomsQuality 5

SquareFootageGarden 5

PreviousOwnerRating 5

HeatingCosts 484

WindowModelNames 5

Price 5

=> Sollen alle Werte entfernt werden (oder nur wo in mehr als z.b. 2 Spalten etwas fehlt)? Gemittelt?

After removal of the 5 cases off all missing values:

Bedrooms 377

Bathrooms 365

SquareFootageHouse 0

Location 194

Age 127

PoolQuality 643

HasPhotovoltaics 79

HeatingType 0

HasFiberglass 0

IsFurnished 0

DateSinceForSale 0

HouseColor 8

PreviousOwnerName 146

HasFireplace 0

KitchensQuality 0

BathroomsQuality 0

BedroomsQuality 0

LivingRoomsQuality 0

SquareFootageGarden 0

PreviousOwnerRating 0

HeatingCosts 479

WindowModelNames 0

Price 0