# Modelling Competition: House prices

## House pricing models

The characteristics of a house strongly affect the sale price. The direction of the effect of most of these characteristics is obvious: larger houses sell for more than smaller houses, houses with more bedrooms and bathrooms sell for more than houses with fewer bedrooms than bathrooms, the presence of a garage tends to raise the sale price of a house, and so on. On the other hand, these features by no means provide a perfect prediction of the sale price of a house. In part, this is because some features are not recorded systematically by real estate agents. For example, whether there is a busy and noisy street in front of the house will matter, whether the house has been kept in good repair is important, and so on. Location usually quite important. For example, the location of the house determines what public school any children living there will attend, and some schools are regarded as better than other schools. Even if there were a complete list of all the features there still would not be a perfect description of the price, because the price is also affected by who buys the house. Houses and buyers are all different. If a family looks at a house that has just been advertised for sale and that house is "just right" for them, it may well sell for the list price or even a little higher. On the other hand, if the house has been for sale for a long time and a family sees that the house will meet their needs only with some changes then it may well sell for quite a bit less than the list price. Sending someone around to look at each house each year and estimate its value is expensive. It is much cheaper to look at the record of the characteristics of each house, and then use a pricing model to estimate the value of the house

## Aim

1. Create an effective price prediction model
2. Identify the important home price attributes
3. Validate the model's prediction accuracy

## Data

The data set collects every aspect of almost 3000 residential homes in the US, as well as their sale price. This are variables:

- **SalePrice**: the property's sale price in dollars. This is the target variable that you're trying to predict.
- **MSSubClass**: The building class
- **MSZoning**: The general zoning classification
- **LotFrontage**: Linear feet of street connected to property
- **LotArea**: Lot size in square feet
- **Street**: Type of road access
- **Alley**: Type of alley access
- **LotShape**: General shape of property
- **LandContour**: Flatness of the property
- **Utilities**: Type of utilities available
- **LotConfig**: Lot configuration
- **LandSlope**: Slope of property
- **Neighborhood**: Physical locations within Ames city limits
- **Condition1**: Proximity to main road or railroad
- **Condition2**: Proximity to main road or railroad (if a second is present)

- **BldgType**: Type of dwelling
- **HouseStyle**: Style of dwelling
- **OverallQual**: Overall material and finish quality
- **OverallCond**: Overall condition rating
- **YearBuilt**: Original construction date
- **YearRemodAdd**: Remodel date
- **RoofStyle**: Type of roof
- **RoofMatl**: Roof material
- **Exterior1st**: Exterior covering on house
- **Exterior2nd**: Exterior covering on house (if more than one material)
- **MasVnrType**: Masonry veneer type
- **MasVnrArea**: Masonry veneer area in square feet
- **ExterQual**: Exterior material quality
- **ExterCond**: Present condition of the material on the exterior
- **Foundation**: Type of foundation
- **BsmtQual**: Height of the basement
- **BsmtCond**: General condition of the basement
- **BsmtExposure**: Walkout or garden level basement walls
- **BsmtFinType1**: Quality of basement finished area
- **BsmtFinSF1**: Type 1 finished square feet
- **BsmtFinType2**: Quality of second finished area (if present)
- **BsmtFinSF2**: Type 2 finished square feet
- **BsmtUnfSF**: Unfinished square feet of basement area
- **TotalBsmtSF**: Total square feet of basement area
- **Heating**: Type of heating
- **HeatingQC**: Heating quality and condition
- **CentralAir**: Central air conditioning
- **Electrical**: Electrical system
- **1stFlrSF**: First Floor square feet
- **2ndFlrSF**: Second floor square feet
- **LowQualFinSF**: Low quality finished square feet (all floors)
- **GrLivArea**: Above grade (ground) living area square feet
- **BsmtFullBath**: Basement full bathrooms
- **BsmtHalfBath**: Basement half bathrooms
- **FullBath**: Full bathrooms above grade
- **HalfBath**: Half baths above grade
- **Bedroom**: Number of bedrooms above basement level
- **Kitchen**: Number of kitchens
- **KitchenQual**: Kitchen quality
- **TotRmsAbvGrd**: Total rooms above grade (does not include bathrooms)
- **Functional**: Home functionality rating
- **Fireplaces**: Number of fireplaces
- **FireplaceQu**: Fireplace quality
- **GarageType**: Garage location
- **GarageYrBlt**: Year garage was built
- **GarageFinish**: Interior finish of the garage
- **GarageCars**: Size of garage in car capacity
- **GarageArea**: Size of garage in square feet
- **GarageQual**: Garage quality
- **GarageCond**: Garage condition
- **PavedDrive**: Paved driveway
- **WoodDeckSF**: Wood deck area in square feet
- **OpenPorchSF**: Open porch area in square feet
- **EnclosedPorch**: Enclosed porch area in square feet

- **3SsnPorch**: Three season porch area in square feet
- **ScreenPorch**: Screen porch area in square feet
- **PoolArea**: Pool area in square feet
- **PoolQC**: Pool quality
- **Fence**: Fence quality
- **MiscFeature**: Miscellaneous feature not covered in other categories
- **MiscVal**: Value of miscellaneous feature
- **MoSold**: Month Sold
- **YrSold**: Year Sold
- **SaleType**: Type of sale
- **SaleCondition**: Condition of sale

### Files

Four files are provided:

- data_description.txt: full description of each column in the dataset
- train.csv: the training set
- test.csv: the test set
- predicted_prices.xlsx (where you have to include the predictions from the test dataset)

## Task

It is your job to predict the sales price for each house. For each Id in the test set, you must predict the value of the SalePrice variable (**work with the log of the Saleprice**). There are many models that could be set up, but you can use **only the techniques studied in this course**.

You must return a report (no longer than 15 pages) in .pdf containing:

1. Model-building steps (exploratory analysis, model fitting, variable selection, etc.).
2. Best model/models selected (There are more than one good model, so give reasons to justify your election).
3. Goodness of fit of the model within sample and predictive power of the model out of sample (in terms of the RMSE)
4. Other questions to be assessed:
    - What is the variable that contributes the most to explain the variability in price?, what are the characteristics of the cheapest/most expensive houses?
    - Any other question that you consider interesting

## Evaluation

1. Clarity: The report must clearly indicate the steps you follow in your work, including which models are considered. Being able to communicate effectively is extremely important for a researcher. Excellent work is of no value if no other than the investigator can understand it. (20%)
2. Content: The analysis should use the tools developed in the course in an appropriate and correct manner. The report should anticipate questions that a critical reader might ask. (60%)
3. Presentation: The results will be presented in the last day of the course. The exposition and defense of the work will be evaluated (20%)

## Winners of the competition

- The group reporting the smallest RMSE will have 1 extra points in the final mark of the course. The runner-up group 0.5