1. Problem Statement

How to give a price range to home buyer, by predicting the price, based on the different features of a house, including no of bedrooms, size of the living area, etc.…

1. Data set Source

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>

1. Data set Details

The data has 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, USA.

Records details

**Total No of Rows: 1460**

**Total No of Columns: 80**

## Here's a brief version of what you'll find in the data description file.

* 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

1. Initial findings
   1. Some of the features, like Alley and PoolQC have no observations data, and due to many features to be analyzed and considered for the prediction, we forced to remove the features with 30% or less NaN values.
   2. The total no of features are divided into the following

Total No of Quantitative Features: 37

Total No of Qualitative Features: 39

* 1. The Sale price distribution details are

Min Selling Price is: 34900

Average Selling Price is: 180921.20

Median Selling Price is: 163000.0

Max Selling Price is: 755000

* 1. It is apparent that Sale Price doesn't follow normal distribution, so before performing regression it has to be transformed. While log transformation does pretty good job, best fit is unbounded Johnson distribution.
  2. When the data divided into multiples category by range of Sale Prices, like 100000, 200000 and 300000 is considered to be low, mid and high range. We found that more houses falling in the mid-range as compared to low and high
  3. Some categories seem to more diverse with respect to Sale Price than others. Neighborhood has big impact on house prices. Most expensive seems to be Partial Sale Condition. Having pool on property seems to improve price substantially. There are also differences in variabilities between category values.

1. Relationships Significant

|  |  |
| --- | --- |
| **Feature Name** | **Percentage of Relationships Significant** |
| **MSSubClass** | 97.70% |
| **GarageType** | 93.00% |
| **Neighborhood** | 90.70% |
| **ExterQual** | 90.70% |
| **BldgType** | 90.70% |
| **Exterior2nd** | 90.70% |
| **BsmtQual** | 88.40% |
| **Exterior1st** | 88.40% |
| **KitchenQual** | 88.40% |
| **HouseStyle** | 88.40% |
| **BsmtFinType1** | 86.00% |
| **BsmtExposure** | 86.00% |
| **MSZoning** | 83.70% |
| **Foundation** | 83.70% |
| **BsmtFinType2** | 83.70% |
| **FireplaceQu** | 81.40% |
| **GarageCond** | 81.40% |
| **GarageFinish** | 81.40% |
| **PavedDrive** | 81.40% |
| **SaleCondition** | 81.40% |
| **HeatingQC** | 79.10% |
| **CentralAir** | 76.70% |
| **MasVnrType** | 76.70% |
| **GarageQual** | 76.70% |
| **BsmtCond** | 76.70% |
| **LotShape** | 74.40% |
| **RoofStyle** | 74.40% |
| **Electrical** | 72.10% |
| **Alley** | 69.80% |
| **LandContour** | 69.80% |
| **ExterCond** | 69.80% |
| **Functional** | 67.40% |
| **SaleType** | 65.10% |
| **Condition1** | 60.50% |
| **Heating** | 58.10% |
| **Fence** | 55.80% |
| **LotConfig** | 46.50% |
| **RoofMatl** | 46.50% |
| **LandSlope** | 39.50% |
| **Condition2** | 39.50% |
| **MiscFeature** | 37.20% |
| **Street** | 37.20% |
| **PoolQC** | 34.90% |
| **Utilities** | 20.90% |