University of Cyprus

# Ames Housing Dataset Analysis

A DSC-510 Introduction to Data Science and Analytics group project

Authors:

- Marios Theofanous
- Antonis Alexandrou
- Ioannis Dimitriou

## Introduction

The housing market is a key domain within real estate and economics, where understanding property values plays an important role for buyers, sellers, investors, and policymakers. A complex mix of factors are influencing the prices, for example location, physical characteristics, neighborhood amenities, and economic conditions. Accurately estimating a property's value can help stakeholders make informed decisions, optimize investments, and improve affordability analyses.

The Ames Housing dataset provides a rich and realistic representation of this domain, containing detailed information on *79 features* describing 2930 residential properties sold in Ames, Iowa. The specific problem is to predict the sale price of a house based on these features. This involves addressing challenges such as handling missing data and encoding categorical variables.

The goal of the project is to build a predictive model that accurately estimates house sale prices using the given features. Beyond simple prediction, the objective is to explore which factors most strongly influence property values using methodologies taught in the course and evaluate their performance. A well-performing model could offer insights into the housing market, assist in automated valuation systems, and demonstrate key machine learning techniques in a real-world context.

# Dataset Description

The Ames Housing dataset was compiled by De Cock (2011) as a modern alternative to the classic Boston Housing dataset. It contains information on residential properties sold in Ames, Iowa, between 2006 and 2010. The data was collected from public property records, including details provided by the local assessor's office and other official sources, capturing a wide range of property characteristics. This makes the dataset a realistic representation of the housing market in a mid-sized U.S. city.

The dataset contains 79 explanatory variables describing various aspects of each property. These include numerical features such as lot area, total basement area, and year built, as well as categorical and ordinal features such as neighborhood, building type, garage quality, and overall condition. Many features capture structural details, amenities, and location factors that influence house prices. Some variables have missing values, requiring preprocessing before analysis.

The target variable for this dataset is `SalePrice`, representing the final sale price of each house in U.S. dollars. The main objective of a predictive modeling task with this dataset is to accurately estimate `SalePrice` using the provided features, making it a supervised regression problem.

## Key Features in the Ames Housing Dataset

Here we present what we consider to be the five most representative features. The overall effect these features have on predicting the target variable will be discussed in the next report, along with the exploratory data analysis:

**1. `OverallQual` (Overall Material and Finish Quality)**

This is an ordinal variable on a 1–10 scale rating the overall quality of the house's construction and finish. It is one of the most influential predictors of housing sale price, as higher-quality homes typically command higher sale prices.

**2. `GrLivArea` (Above Ground Living Area)**

`GrLivArea` measures the total above-ground living space in square feet. Since larger homes sell for higher prices in general, this variable provides a direct measure of how house size influences value.

**3. `Neighborhood`**

`Neighborhood` is a categorical variable indicating the physical location of the house within Ames, Iowa. It captures differences in safety, amenities, and overall desirability of the area. Location consistently ranks as a critical factor in real estate pricing, and homes in sought-after neighborhoods often have much higher sale prices.

**4. `YearBuilt`**
`YearBuilt` is a numeric variable that records the year in which the house was constructed. Newer homes command higher sale prices due to modern construction standards and energy-efficient designs. However, older homes located in historic or prestigious neighborhoods can also achieve premium prices, making this variable an interesting balance between age and desirability.

5. **`GarageCars / GarageArea`**
These numeric variables represent the number of cars a garage can accommodate and the total garage area in square feet. Garages add both practical utility and desirability to a property. Homes with larger or multiple garages tend to be priced higher.

# References

- De Cock, D. (2011). Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. Journal of Statistics Education, 19(3).

  Link: https://jse.amstat.org/v19n3/decock.pdf (accessed 2025)

- Wang, Ziwen (2024). Research on predicting Ames housing price based on forward selection regression and principal component regression. Applied and Computational Engineering.

  Link:
  https://www.researchgate.net/publication/382373765_Research_on_predicting_Ames_housing_price_based_on_forward_selection_regression_and_principal_component_regression

- A study on Regression applied to the Ames dataset: This comprehensive study explores various regression techniques, offering a deep dive into model selection and evaluation metrics.

  Link: https://www.kaggle.com/code/juliencs/a-study-on-regression-applied-to-the-ames-dataset