

**STATISTICAL INFERENCE AND MODELING**

**ASSIGNMENT 1: HOUSE PRICES**

*November 2023*

Tutor: Lidia Montero Mercadé

Adrià Casanova

Víctor Garcia

Zhengyong Ji

Index

[Data Pre-processing 3](#_Toc150869257)

[Data Overview 3](#_Toc150869258)

[Data Transformation 3](#_Toc150869259)

[Univariate Outlier 3](#_Toc150869260)

[Multivariate Outlier 3](#_Toc150869261)

[Data Imputation 3](#_Toc150869262)

[Feature selection 3](#_Toc150869263)

[Model Fitting 4](#_Toc150869264)

[Model Validation 5](#_Toc150869265)

[Appendix 6](#_Toc150869266)

# Data Pre-processing

In this work, we’ll study the data set called “Ames Housing dataset”, collected by Dean De Cock for the purpose to analyze the correlation about house prices and different features that describe the house condition, and then to build a regression model that will allows us to predict the sale price.

## Data Overview

The data set has two parts, the training part and testing part, with 1460 and 1459 observations each other, and 81 variables [[1]](#footnote-1)(including the id variable).

## Data Transformation

As the first step, we’ll transform the data set

## Univariate Outlier

In order to detect the univariate outlier, we implement a function that analyze the boxplot of each feature of test dataset.

## Multivariate Outlier

In order to detect multivariate outlier, we used the function Moutlier() from Chemometrics library. As this is a stochastic model, we had to determine manually the feature that we want to impute.

(LotFrontage, LotArea, YearBuilt, BsmtFinSF1, Fireplaces, OpenPorchSF, YrSold)

## Data Imputation

In this section, we’ll impute the previous data set with imputePCA function.

## Feature selection

# Model Fitting

After the preprocessing of data is finished, we will start with the building of our regression model using the numerical feature, and once the model is well fitted, then we’ll add the categorical feature on by one. For each step, we have checked if the model is robust or not, and their multicollinearity.

# Model Validation

# Appendix

1. The target variable, which is the price of the house (SalePrice), is empty for test data set. [↑](#footnote-ref-1)