Project title, description, and goals

- Github repo: https://github.com/matthewgraca/4661-house-prices-regression

The project title is "House Prices Regression", where the goal is to predict the sale price of a house, given several features. The data can be found here: https://www.kaggle.com/c/house-prices-advanced-regression-techniques. The goal is to minimize the root mean squared error of our chosen model with the use of various technqiues:

- Data visualiztion and feature selection.
- Feature engineering with encoders, imputation.
- Dimensionality reduction with PCA.
- Factor Analysis of Multiple Data.
- Hyperparameter tuning with different datasets, models (linear, polynomial), and model parameters (regularization).

Team member responsibilities

Matthew Graca

- Project lead
- Feature engineering
- Dimensionality reduction: PCA, FAMD

George Melendrez

- Hyperparameter tuning

Yash Harishkumar Patel

- Data visualization, correlation analysis

Reda Masri

- Basic model training
- Model analysis, feature selection

Information about the data

Our data has 1460 samples containing 81 columns; that is, 80 features and one target column, 'SalePrice'.

In regards to the completeness of the data, there are 19 columns containing NaN values, that will have to either be removed as a feature or imputed with a value.

In terms of the type of the data, there are 28 columns with trainable data (which excludes useless columns like 'Id' and the target column). There are 13 columns

which can be considered ordinal by nature, and 38 columns that are categorical but cannot be considered ordinal.

Project status

So far, the project has made substantial progress.

We have done data visualization using pair plots to identify correlated features, and have created a correlation matrix for more precise analysis.

We have trained a basic model that will be used as the baseline for future improvements over the course of the project, as we finalize the kind of data that will be used.

We have created mulitple datasets for testing:

- Numerical dataset
- Numerical dataset with one hot encoded categorical features
- Dataset treated with FAMD

We have also implemented PCA for dimensionality reduction.

Finally, we've tested a few hyperparameters with a few different regression models.