PCA with Housing Dataset

Context

Consider the situation where you are working for Zillow as a data scientist.

Housing pricing predictions is the goal.

We know 80 things about each house to use as inputs to be able to predict the price of a house.

Your goal is to islotate the important features from the dataset and build a model which can be used to predict the price of the houses. '

Since there are too many features, Principal Component Analysis can be applied to reduce the number of features used for the actual prediction model, without any loss of information,

Dataset to Work With

https://www.kaggle.com/c/home-data-for-ml-course/data?select=train.csv

- Download the **train.csv** file and load it using pandas

Brief Description of Dataset

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

Detailed Description of Dataset
https://www.kaggle.com/c/home-data-for-ml-course/data?select=data_description.txt

Assignments

Data Cleanup and Exploratory Data Analysis (25%)

1. Explore Basic Statistics of each feature
2. Outlier Detection
    a. pick your own method to remove outliers
    b. Explain the rationale
3. Missing Value Imputations
    a. Use your method of choice for imputation
    b. Explain the rationale
4. Correlation Analysis
    a. Identify the target variable correctly
    b. Isolate the features with high correlation with the target variable

Feature Preparation and Transformation (25%)

1. Drop Unnecessary Columns (All categorical variables, essentially)
2. Apply Scaling to dataset to bring all variables to the same scale
3. Feature Selection for isolating final set of variables for PCA

PCA (25%)

1. Threshold for Variance (90% - industry standard)
2. Balance the number of features selected

Linear Regression (25%)

1. Fit model to cleaned-up dataset
2. Comparative Study of with and without PCA