Project 2:Ames Housing Dataset

Saba Suhail

Objectives

To predict the sale prices of houses for various stakeholders

- Analyze the dataset
- Understand relationships between various features by EDA
- Clean data-imputation,deletion
- Graphs for visual understanding
- Develop regression models

What was done?

- Train.csv(has sale prices)
- test.csv(lacks sale prices)

Distribution of sale prices are log transformed

2050 no null entries

Certain columns altogether dropped such as Alley

Comparison with data dictionary

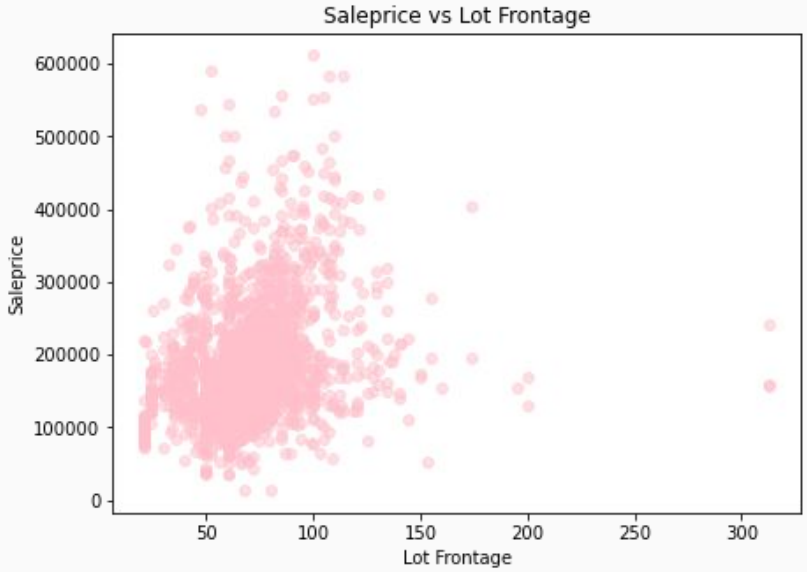
Exhaustive cleaning of train.csv-was it too much?

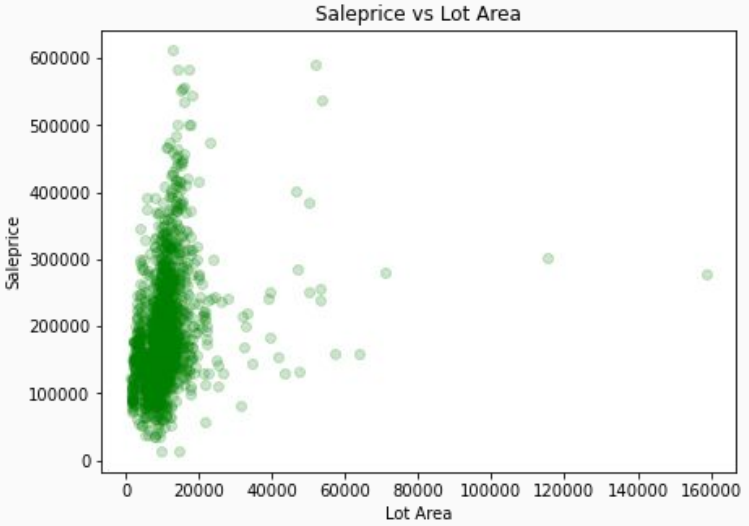
Can build any number of models in the future

Cleaning of test.csv as per requirement

Scatter plot for all features

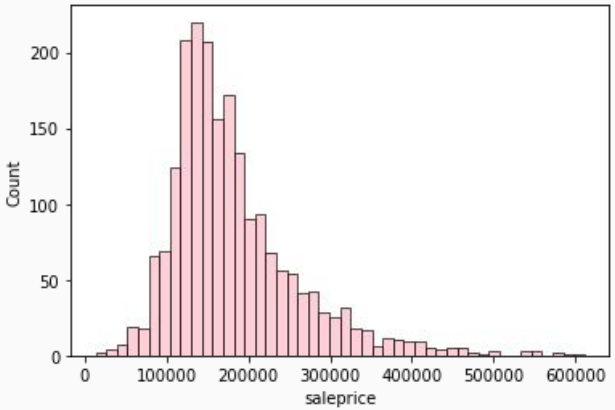
Illustrations



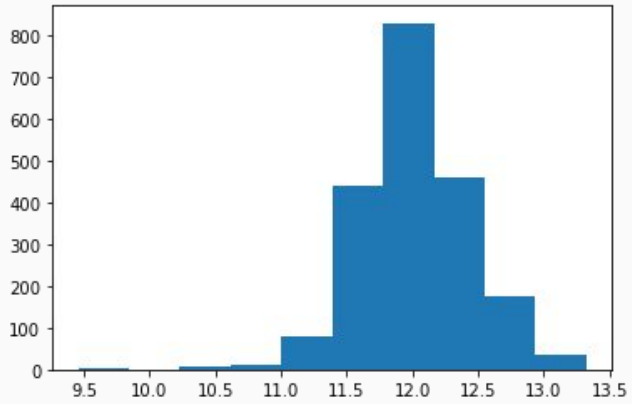


Distribution of sale prices

Without log transformation



With log transformation



Feature Engineered Model

Features Used

- continuous = ['lot_frontage','lot_area','mas_vnr_area','area_sf']
- discrete = ['year_built', 'tot_bath','bsmt_bath','abv_grd_combined']
- nominal= ['ms_subclass', 'ms_zoning','neighborhood','land_contour','central_air']
- ordinal = ['overall_qual','heating_qc']

Features Generated:

- df['tot_bath']=df.full_bath+0.5*df.half_bath
- df['bsmt_bath']=df.bsmt_full_bath + 0.5*df.bsmt_half_bath
- df['abv_grd_combined'] = df.totrms_abvgrd+df.kitchen_abvgr
- df['area_sf']= df['1st_flr_sf']+ df.flr2nd_sf - df.gr_liv_area - df.garage_area + df.open_porch_sf + df.total_bsmt_sf

Models Employed

- Feature Engineered
- Linear Regression with continuous variables
- Ridge Regression with continuous variables
- Lasso Regression with continuous variables

Absurd results

1 for test.csv for upper 2 models

And in the order of 10^(-2) for the lower two

Yet the lower two made much more sense

Issues Faced

Standard Scaler malfunctions at times

Saving a file and then reading it back generates unthinkable problems

Overfitting,absurd results

Would appreciate your feedback here

Would myself really acquaint myself how things are happening and inside maths

The man who blames the supreme certainty of mathematics feeds on confusion, and can never silence the contradictions of sophistical sciences which lead to an eternal quackery.

- Leonardo da Vinci -