HAYA TOUMY

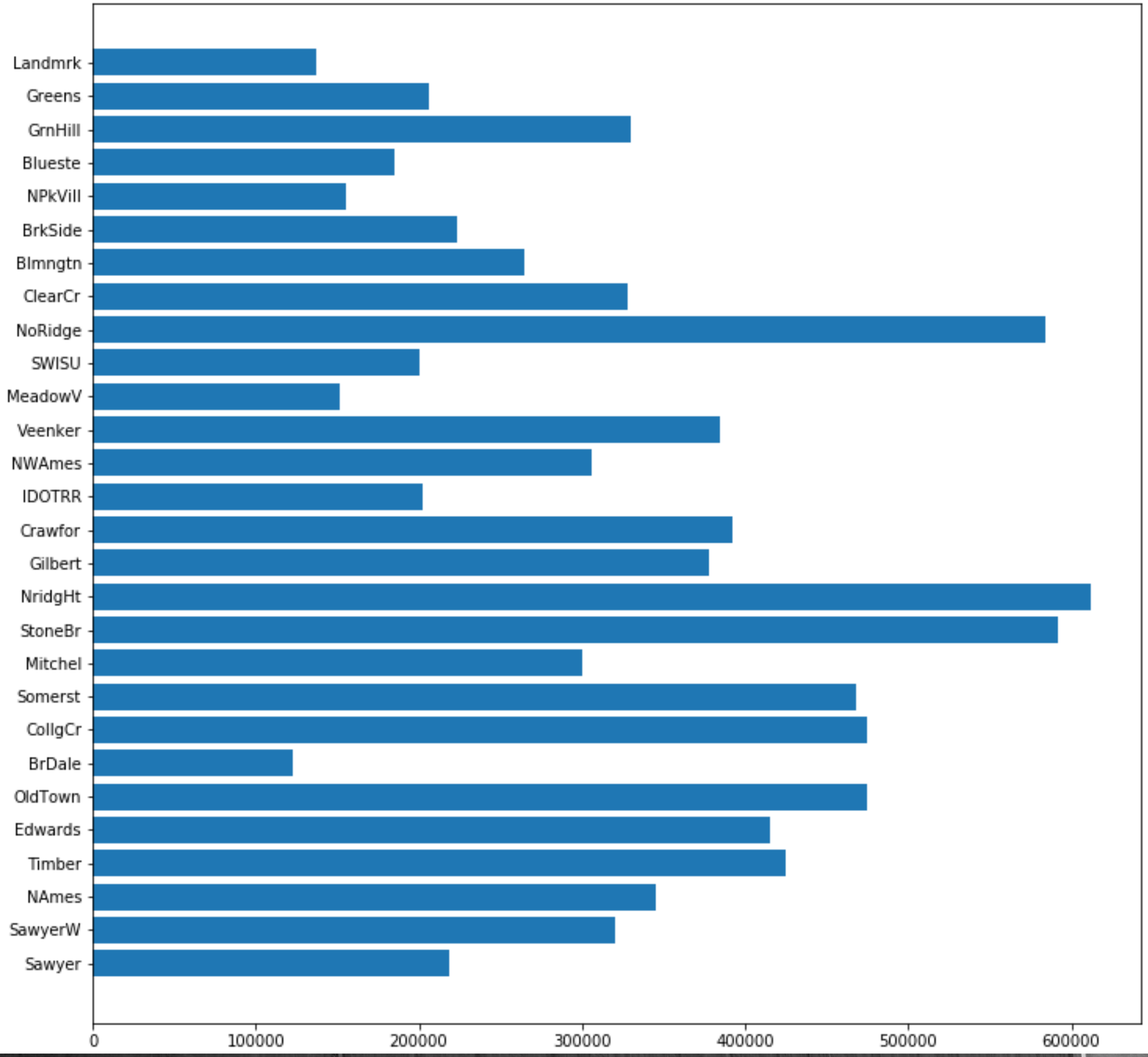
HOUSE SALE PRICES PREDICTIONS. PROJECT_2

AMES, IOWA DATASET FEATURES & TARGET

- 81 variables.
- Split into training set and testing set.
- Many missing values
- Target: predict house selling prices. And give recommendations

FINDINGS IN EXPLORATORY ANALYSIS

- Numerical variables, moderately to strongly correlated with sale prices are:
- Overall quality 0.8 (on a scale from 1: very poor, to 10: very excellent)
- Above ground living area 0.7 (in squared feet)
- Exterior quality 0.66 (1: poor, to 5: excellent)
- Garage area 0.65 (in sq ft)
- Garage cars capacity 0.65 (number of cars a garage can fit)
- Kitchen quality 0.64 (1: poor, to 5: excellent)
- Total basement area 0.63 (in sq ft)
- First floor area 0.62 (in sq ft)

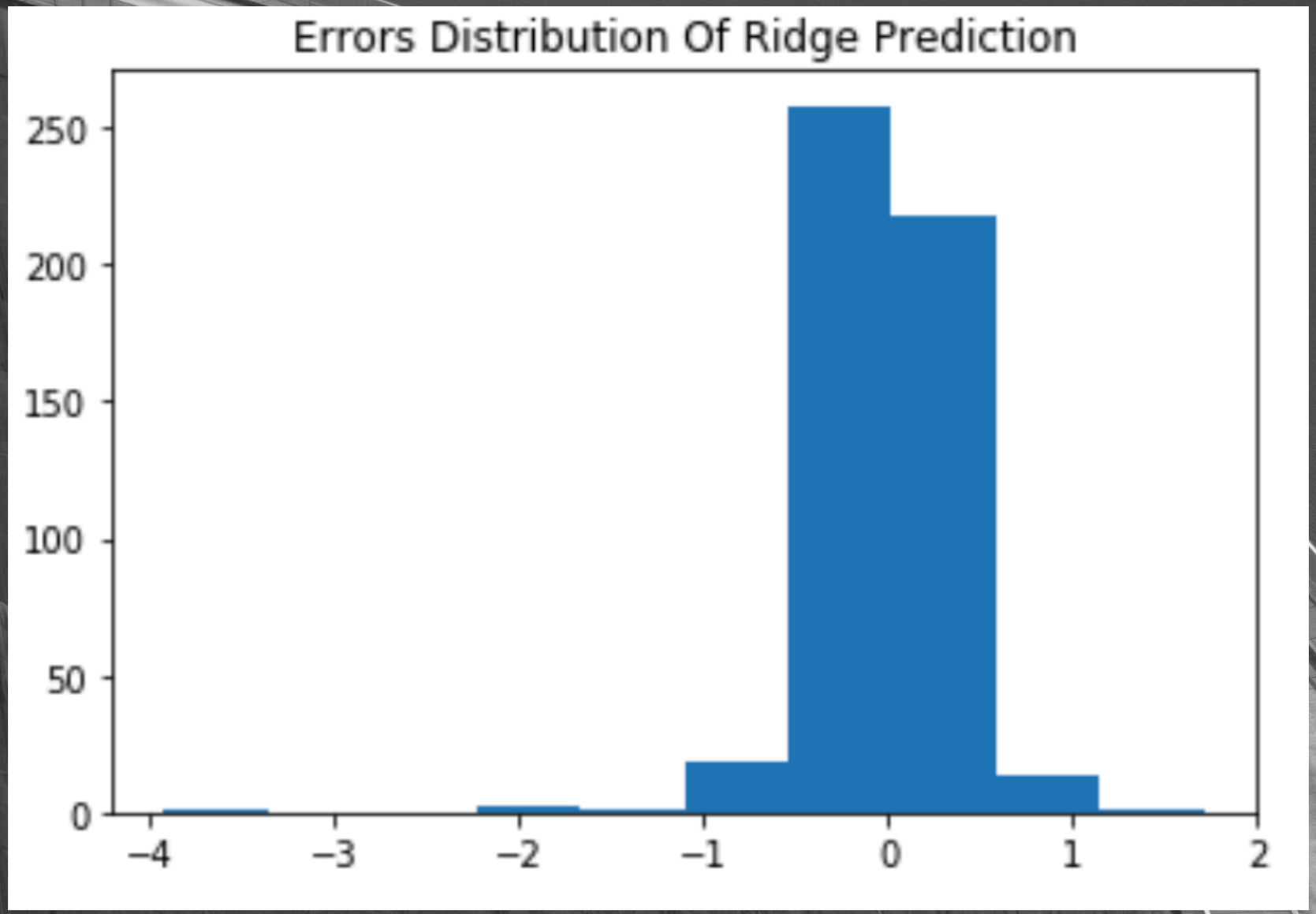

Landmrk
Greens
GrnHill
Blueste
NPkVill
BrkSide
Blmngtn
ClearCr
NoRidge
SWISU
MeadowV
Veenker
NWAmes
IDOTRR
Crawfor
Gilbert
NridgHt
StoneBr
Mitchel
Somerst
CollgCr
BrDale
OldTown
Edwards
Timber
NAmes
SawyerW
Sawyer
0
100000
200000
300000
400000
500000
600000

TEXT

- Top 5 highest selling prices averages, by neighborhood:
- StoneBr $329675.74
- NridgHt $322831.35
- NoRidge $316294.125
- GrnHill $280000
- Veenker $253570.58
- I would say, these are the best neighborhood to invest in.

- The target variable (sale price) is skewed right, suggesting log-transformation to normalize.

Now normalized, as much as possible



Normalizing the target, normalizes predictions, and errors as well. These are the errors of the log-transformed model.

TO EVALUATE MODELS

- Split the training set using 5-folds cross-validation to train the model, and test it.
- I would then found the R2 score to measure goodness of fit to the hold out set.

MODELS ATTEMPTED

- Linear regression: it didn’t perform well, even with Ridge and Lasso.
- Linear regression with polynomial features added to only the strongest correlated variables
- Power transformation, with Ridge, did best with R2 score of 0.894

MOST IMPORTANT FEATURES AND THEIR COEFFICIENTS IN THE MODEL

- Overall quality, and overall condition, range from 10: very excellent, to 1: very poor
- Year built ranges from 1872 to 2010
- Above ground living area ranges between 407 to 4676 sq ft
- Garage car capacity ranged from 0 to 4 cars.

overall_qual	0.321588
year_built	0.210752
gr_liv_area	0.125358
overall_cond	0.117490
garage_cars	0.092161
1st_flr_sf	0.086035
year_remod_add	0.083157
bsmt_full_bath	0.076977
fireplaces	0.069955
2nd_flr_sf	0.067310

RECOMMENDATIONS

- The features appear to add most value are the ones highly positively correlated with SalePrice. The bigger the house, garage, living area, the more expensive the house.
- The better the overall quality of the house, the pricier the house. Also the worse those are, the cheaper the house would sell.
- Neighborhoods for best investment, I'd say are the ones we saw earlier with highest average selling price.

- The model I made can be generalized to other cities, there's no variables I thought specific to Ames, Iowa. That is because I didn't incorporate the neighborhoods in my model.