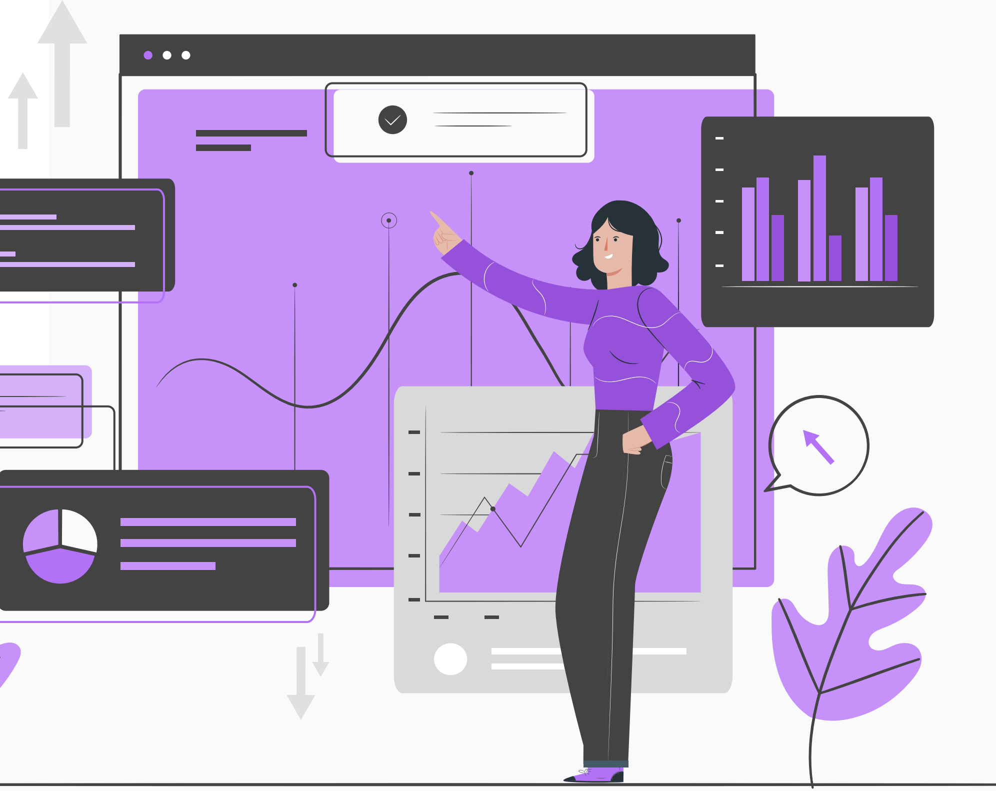
AMES HOUSING PREDICT



DBS-122
Elaine Chen

PURPOSE AND RESOURCE

Purpose: Find the best regression model to predict houses sale price in Ames, IA

Resource:

Data: https://www.kaggle.com/competitions/dsi-122-ames-housing-challenge/data

Google slides template: https://slidesgo.com/

OVERVIEW

01
INTRODUCTION

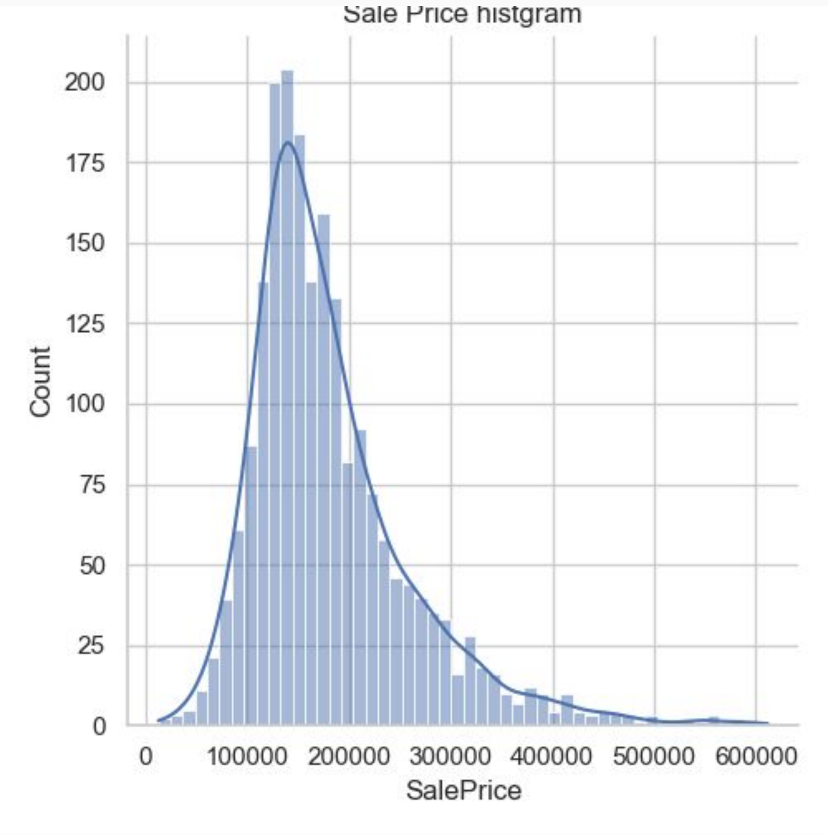
02
DATA CLEANING AND EDA

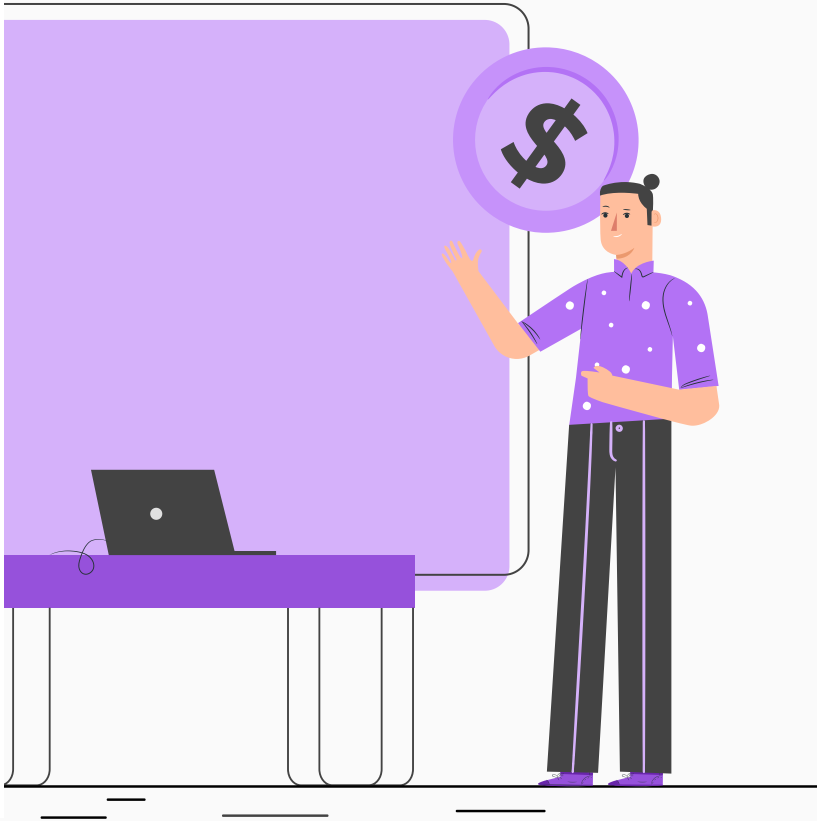
03
TRAINING TIME AND ASSESSMENT

04
CONCLUSIONS

INTRODUCTION

Training data size: 2,051 rows; 81 columns

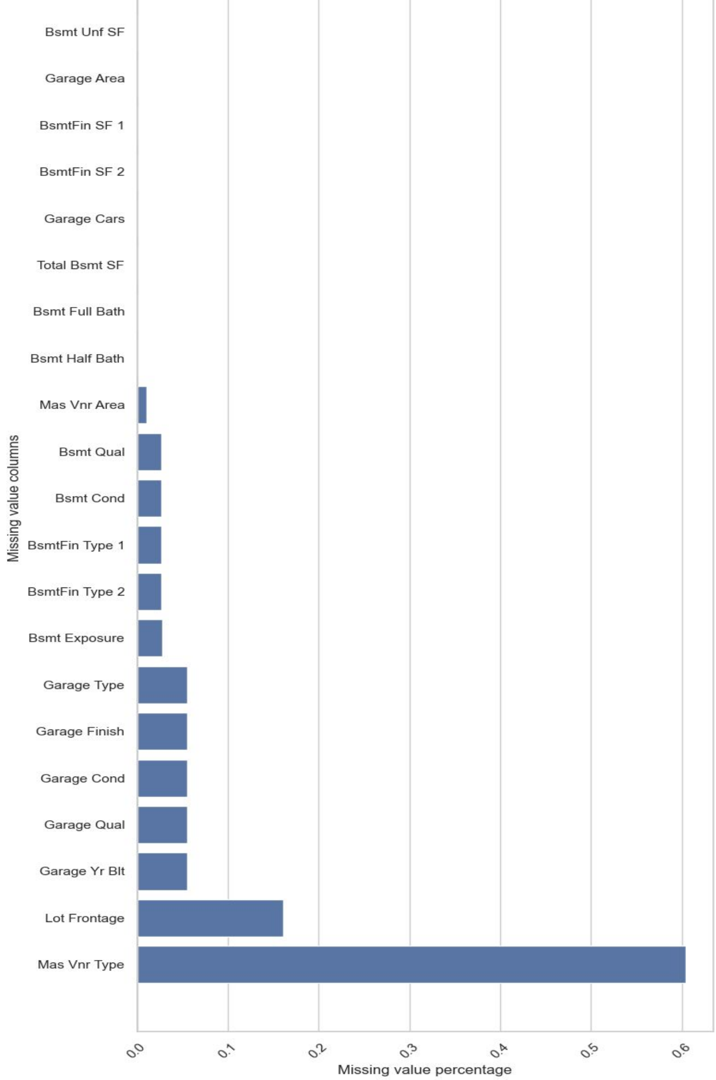




DATA CLEANING AND EDA

Data cleaning:

- Repetitive column: PID
- Missing data
- Outliner



Dropping rows that column missing rate < 0.1%:

- Unfinished basement square feet
- Garage Area
- Type 1 finished square feet
- Quality of second finished area (if present)
- Total basement square feet
- Basement full bathrooms
- Basement half bathrooms

Dropping columns which missing rate>80%

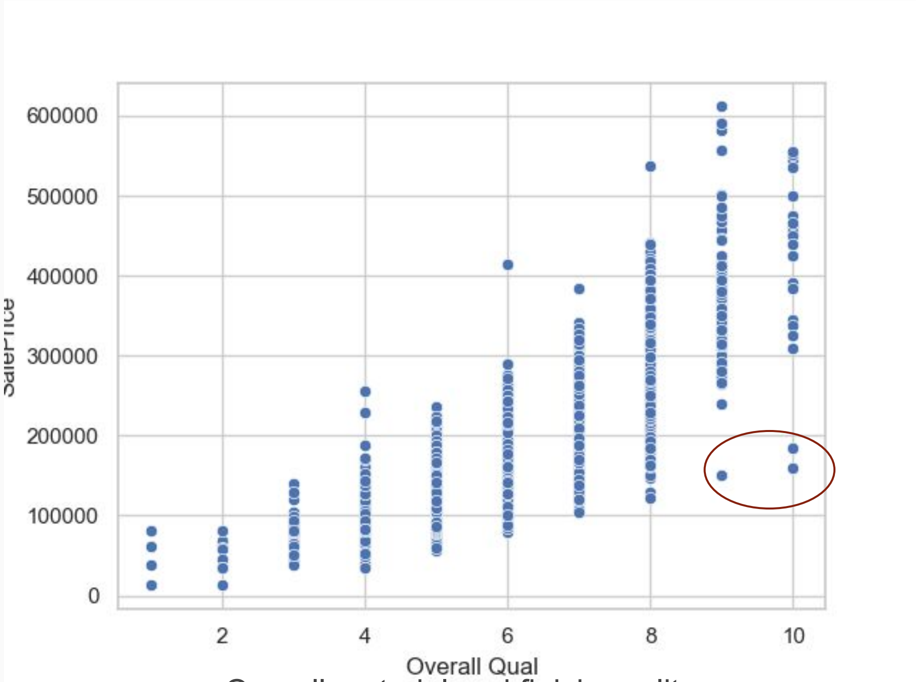
- Fence
- Alley
- Miscellaneous feature not covered in other categories
- Pool quality

Filling/Impute missing value columns

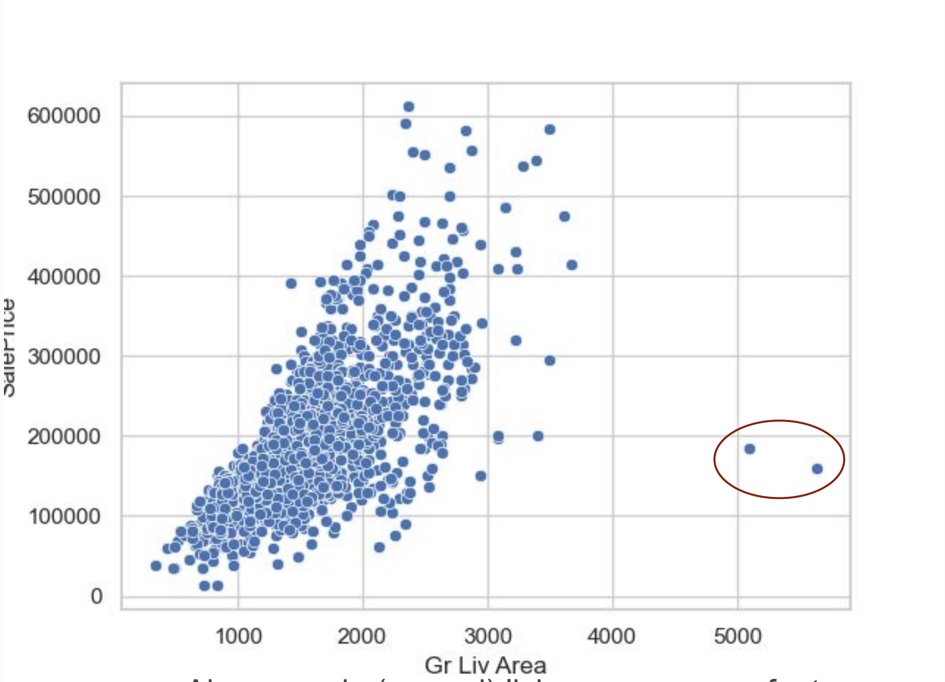
- E.g. Using lot frontage by neighborhood mean to impute missing value of lot frontage

DATA CLEANING AND EDA

Outliers



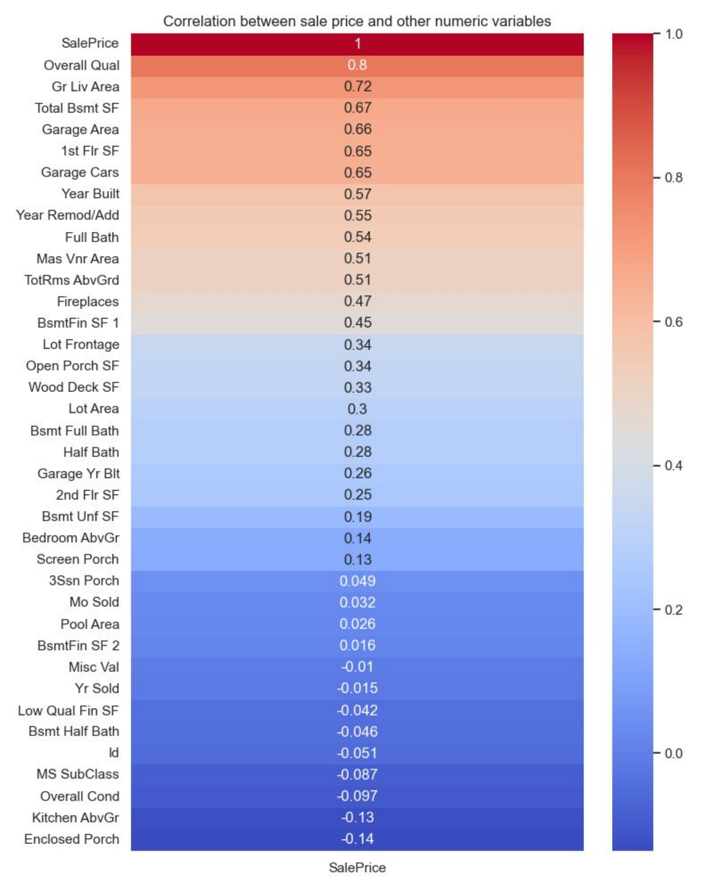
Overall material and finish quality

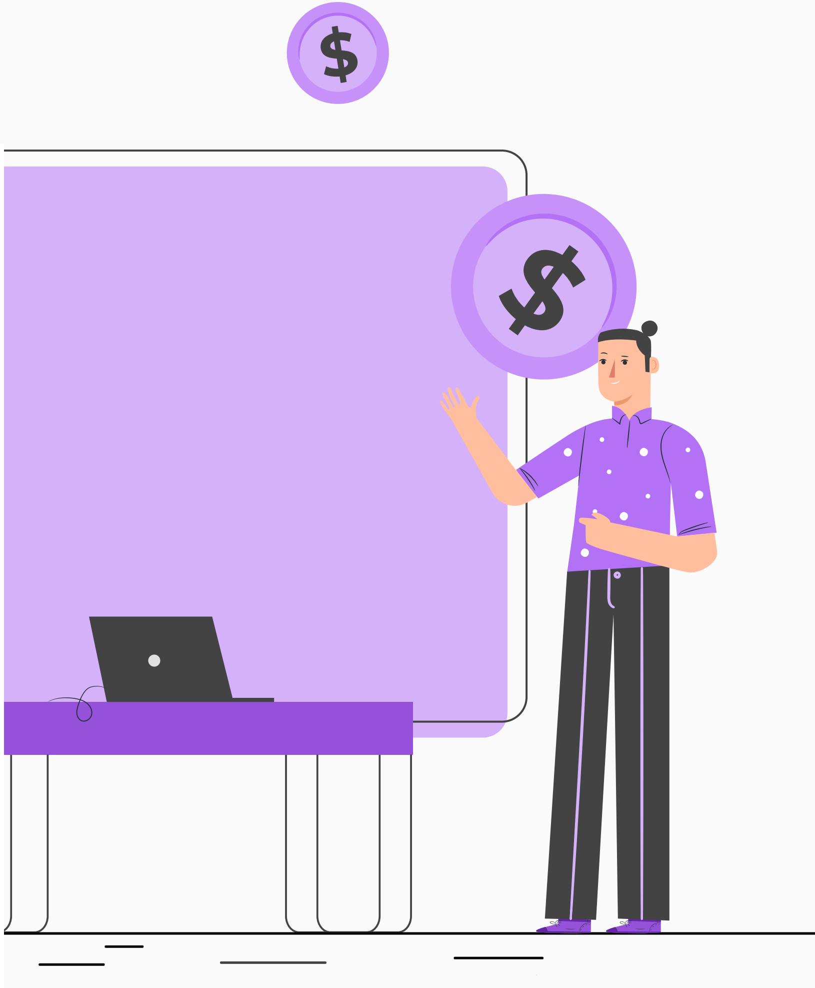


Above grade (ground) living area square feet

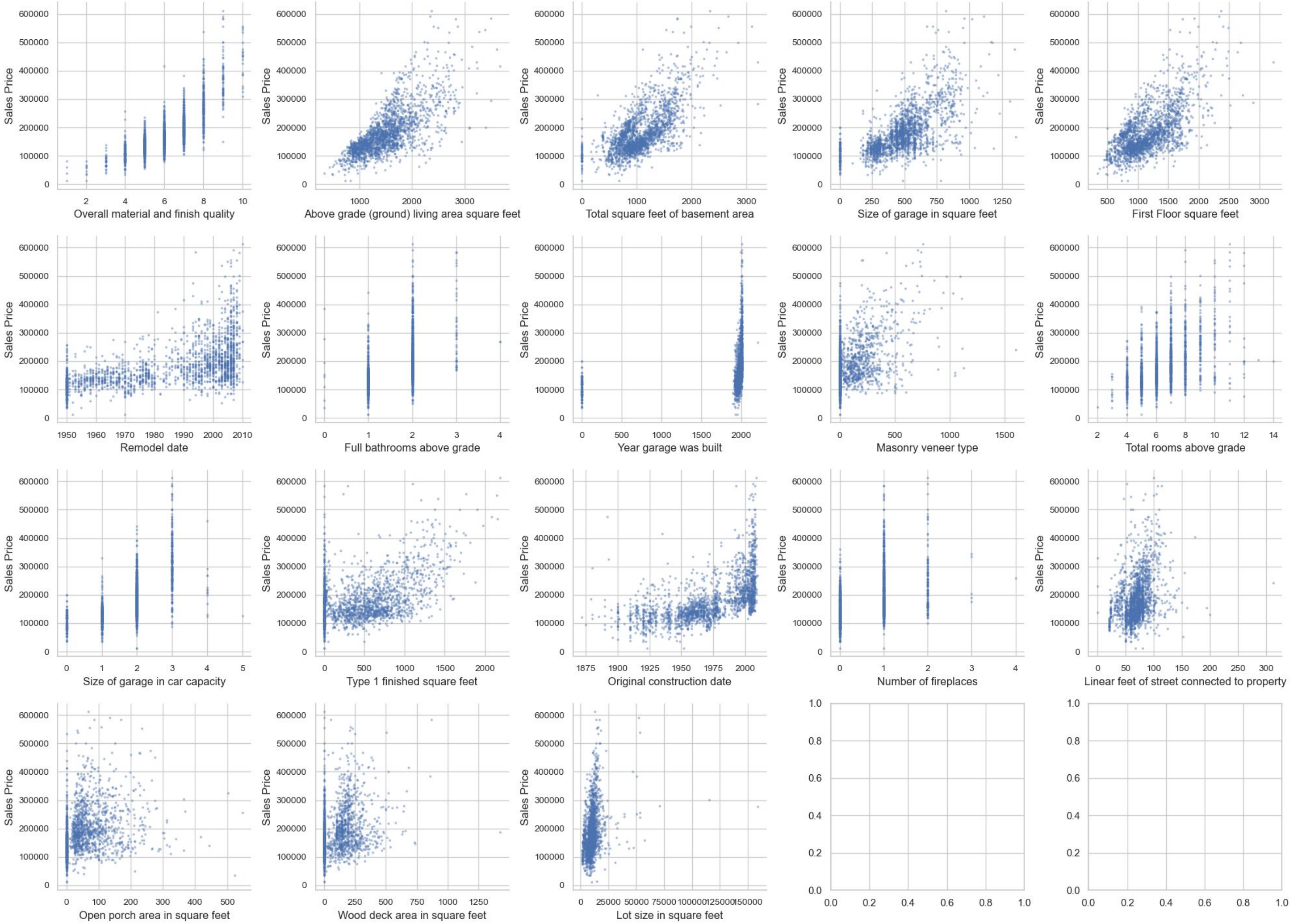
EDA

Select numeric variables that have high correlation with sale price



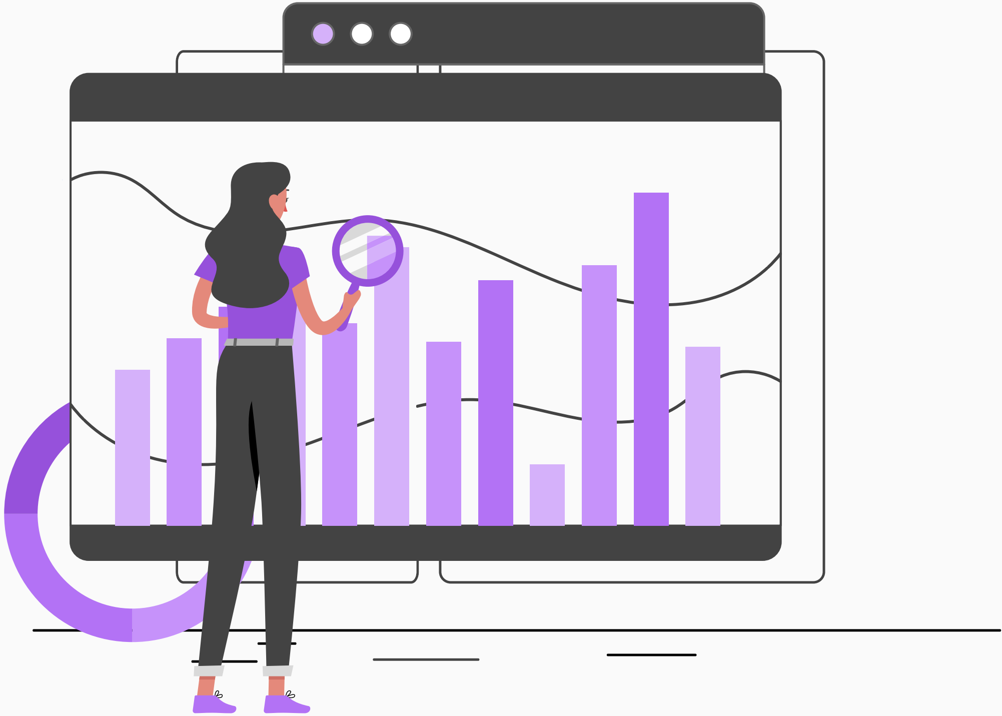


EDA


Sales Price
0
100000
200000
300000
400000
500000
600000
2 4 6 8 10
Overall material and finish quality
1000 2000 3000
Above grade (ground) living area square feet
0 1000 2000 3000
Total square feet of basement area
0 250 500 750 1000 1250
Size of garage in square feet
500 1000 1500 2000 2500 3000
First Floor square feet
1950 1960 1970 1980 1990 2000 2010
Remodel date
0 1 2 3 4
Full bathrooms above grade
0 500 1000 1500 2000
Year garage was built
0 500 1000 1500
Masonry veneer type
2 4 6 8 10 12 14
Total rooms above grade
0 1 2 3 4 5
Size of garage in car capacity
0 500 1000 1500 2000
Type 1 finished square feet
1875 1900 1925 1950 1975 2000
Original construction date
0 1 2 3 4
Number of fireplaces
0 50 100 150 200 250 300
Linear feet of street connected to property
0 100 200 300 400 500
Open porch area in square feet
0 250 500 750 1000 1250
Wood deck area in square feet
0 25000 50000 75000 100000 125000 150000
Lot size in square feet
0.0 0.2 0.4 0.6 0.8 1.0
0.0 0.2 0.4 0.6 0.8 1.0
0.0 0.2 0.4 0.6 0.8 1.0
0.0 0.2 0.4 0.6 0.8 1.0

03.

TRAINING TIME



FEATURES SELECTION

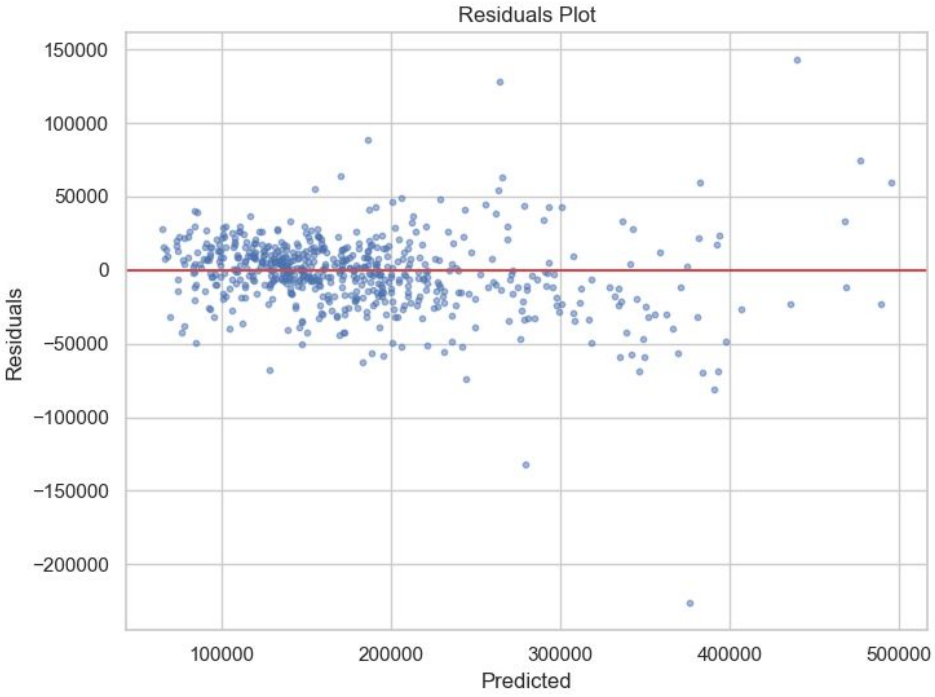
NUMERIC VARIABLES

CATEGORICAL VARIABLES

FEATURES PROCESS

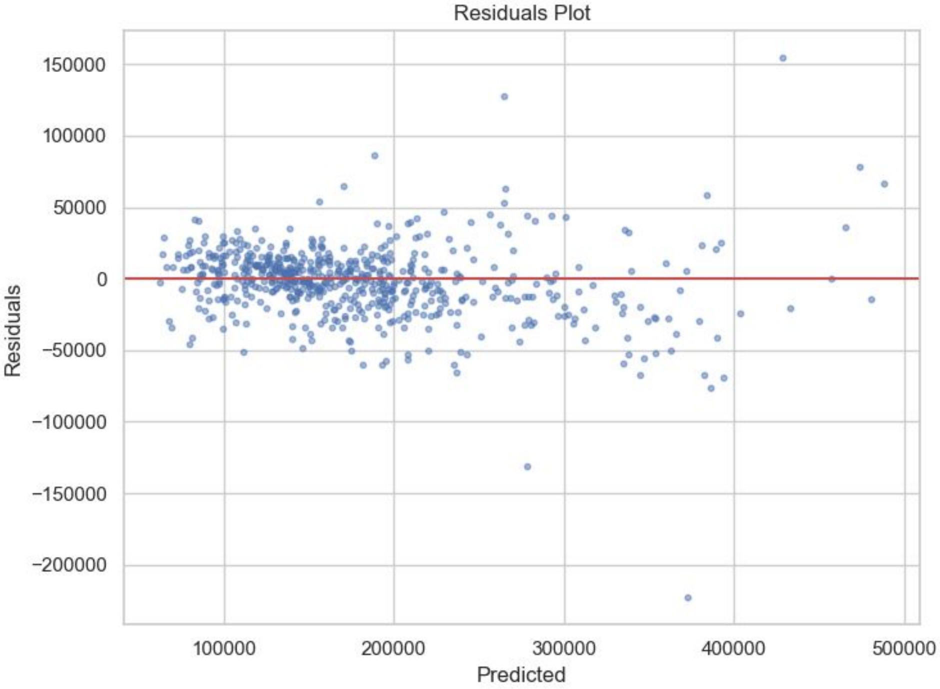
1. Selecting numeric variables having >=0.3 (or <= -0.3) correlation with sale price
2. OneHot Encoded categorical features:
   House Style,
   Neighborhood,
   General zoning classification,
   Kitchen quality,
   Roof material,
   Foundation,
3. Add interaction features:
   Garage Cars,
   Overall quality,
   Fireplaces,
   Full bath,
   Total rooms above grade

OLS COEFFICIENT OF DETERMINATION



RMSE	26130.52
Train model R²	92.99%
Test model R²	88.59%
Mean of cross validation score	86.23%

LASSO COEFFICIENT OF DETERMINATION



RMSE	26010.26
Train model R²	92.85%
Test model R²	88.81%
Mean of cross validation score	87.46%

CONCLUSIONS

In this Ames house sale price model, LASSO regression has better coefficient of determination over Ordinary Least Squares regression. In training dataset, approximately 92.86% of the variance in the sale price can be explained by the features in the LASSO regression model; 88.81% of sale price can be explained by the features in testing dataset.

THANKS

Does anyone have any questions?

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik** and illustrations by **Storyset**