Project 2:

Ames Housing Project

Yang Li

Introduction

- Problem Statement
  - Ames is a town in Iowa with a population of 66,258 in 2019. Ames has highly rated public school and attracts many young professionals to look for house there.

- Objectives
  - As a member of data science team in Ace Real Estate, we will apply machine learning skill to estimate sale price of houses.
  - Build linear regression model to predict the sale price for houses in Ames and provide recommendation for homeowners to increase their house value.

Data Set Description

- The Ames Housing dataset is collected for houses sold between year 2006 to 2010.
- The dataset includes 80 features of nominal, discrete, ordinal and continuous variables for individual residential properties sold.

Data Analysis Processes

- **Data Preparation**
  - Data Cleaning - Missing values were detected and fixed,
  - Outliers Investigation and Elimination
  - Features transformation according to type of variables

- **Features Selection**
  - Features selection with correlation matrix
  - Visualization for selected features
    - Continuous data with scatter plot
    - Discrete data with box plot
  - Check collinearity within features

- **Model Verification**
  - 1st round verification: Apply selected features
  - 2nd round verification: Add power 2 (square) features
  - 3rd round verification: Add power 3 features
  - 4th round verification: Drop features with zero coeffects from 3rd round verification
  - Residual plot with best model

Data Analysis Processes

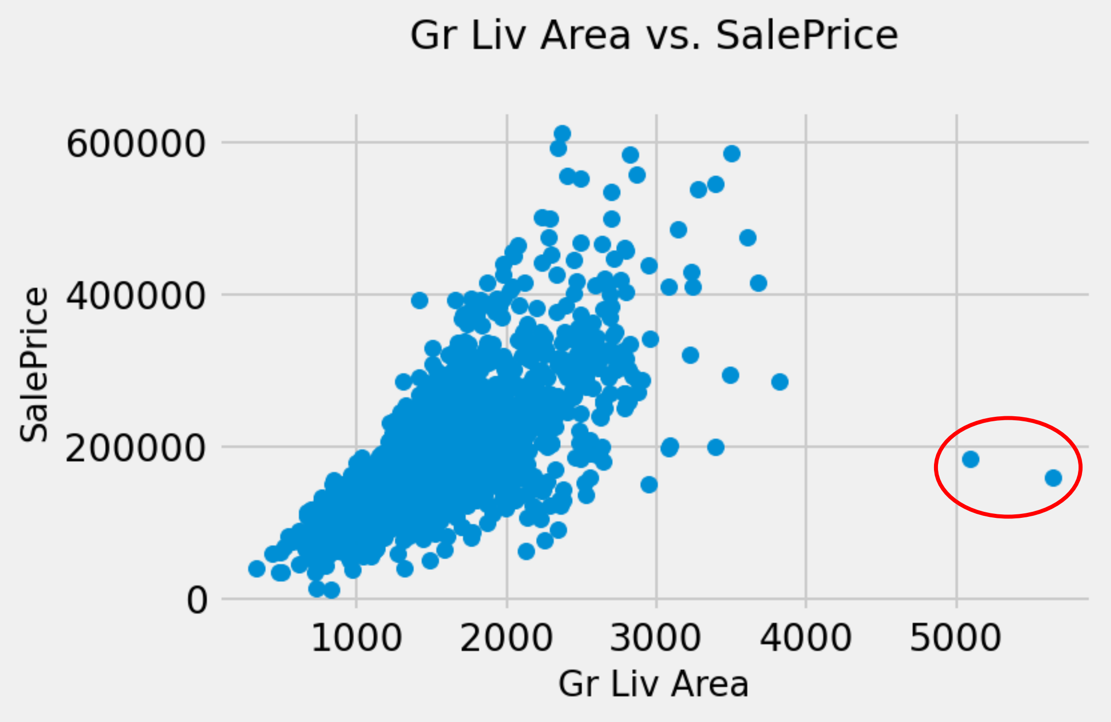
- **Data Preparation - Missing values were detected and fixed,**

Data missing type	Train dataset	Test dataset	Imputation
Most values missing	Alley ---------1911 Pool QC-------2042 Fence---------1651 Misc Feature---1986	Alley ---------821 Pool QC-------875 Fence---------707 Misc Feature---838	Drop these features due to most of values missing
Some values missing (Ordinal or nominal data)	Mas Vnr Type---22 Bsmt Qual------55 Bsmt Cond----- 55 Bsmt Exposure-- 58 BsmtFin Type 1--55 BsmtFin Type 2--1 Fireplace Qu----1000 Garage Type----113 Garage Finish---114 Garage Qual----115 Garage Cond--- 114	Mas Vnr Type---1 Bsmt Qual------25 Bsmt Cond----- 25 Bsmt Exposure-- 25 BsmtFin Type 1--25 BsmtFin Type 2--25 Fireplace Qu----422 Garage Type----44 Garage Finish---45 Garage Qual----45 Garage Cond--- 45	Impute with "None' or 'NA' or 'No' according to data dictionary
Some values missing (Continuous or discrete)	Mas Vnr Area--- 22 BsmtFin SF 1---- 1 BsmtFin SF 2---- 1 Bsmt Unf SF-----1 Total Bsmt SF----1 Bsmt Full Bath--- 2 Bsmt Half Bath---2 Garage Yr Blt---- 114 Garage Cars----- 1 Garage Area-----1	Mas Vnr Area--- 1 Garage Yr Blt---- 45 Electrical-------- 1	Impute with 0.
Some values missing (Continuous)	Lot Frontage ----330	Lot Frontage ----160	Impute with mean for train data and test data separately.

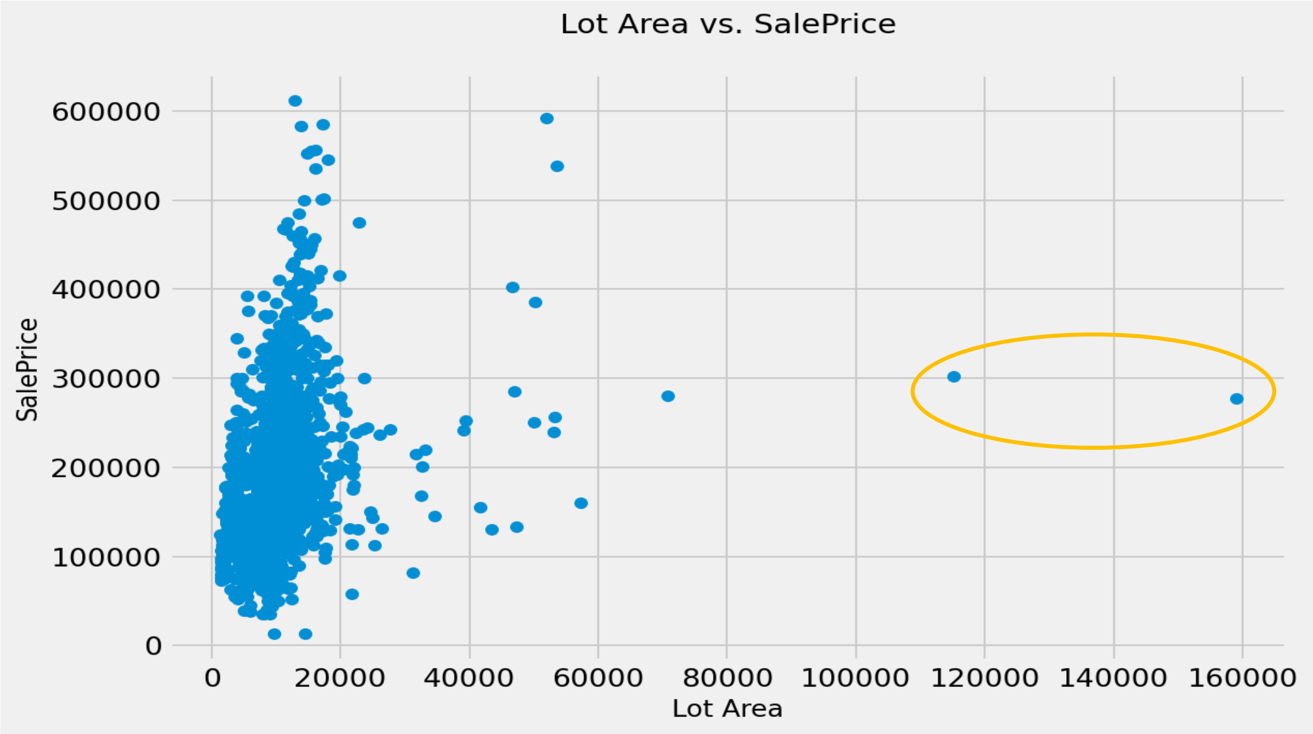
Data Analysis Processes

- **Data Preparation - Outliers Investigation and Elimination**

Outliers for Gr Liv Area > 4000

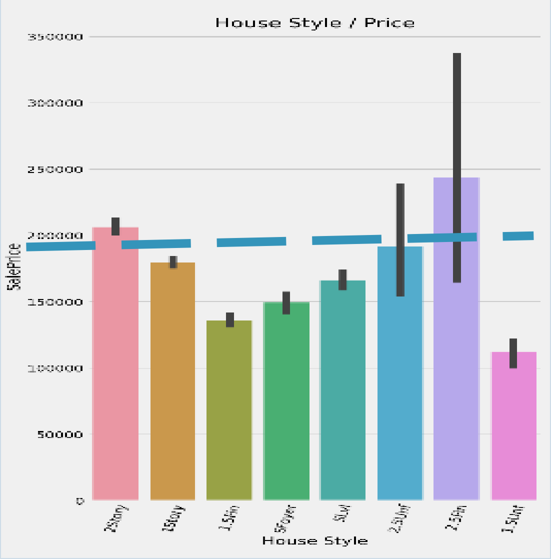


Outliers for Lot Area > 100000



Data Analysis Processes

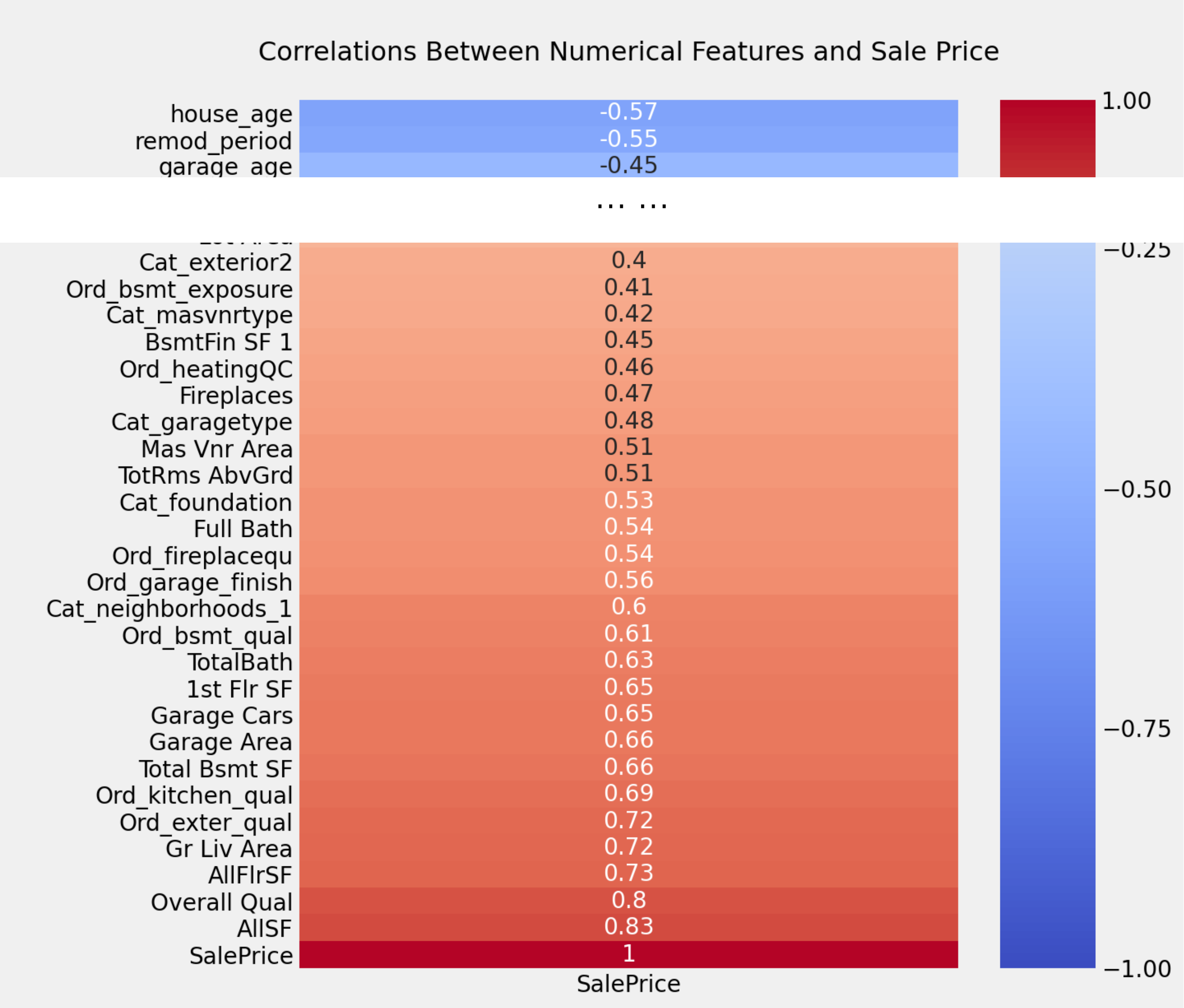
- **Data Preparation - Features transformation according to type of variable**

Type of variable	Transformation	Examples
Continuous	Create new and more meaning features	From ‘Yr sold’, ‘Year Built’, ‘Year Remod/Add’ to create ‘house_age’ & ‘remod_period’, etc
Ordinal Categorical	Manually encode the variable ['Ex’ ,'Gd’ , 'TA’, 'Fa’ , 'Po’] → [5, 4, 3, 2, 1,]	‘Exter Qual’, 'Exter Cond’ etc
Nominal Categorical	According bar charts to group categories which relate to high Sale price. In ‘House Style’, group ‘2Story’ and ‘2.5Fin’.	‘Neighborhood’, ‘House Style etc 

Data Analysis Processes

- **Features Selection - heatmap and correlation matrix**

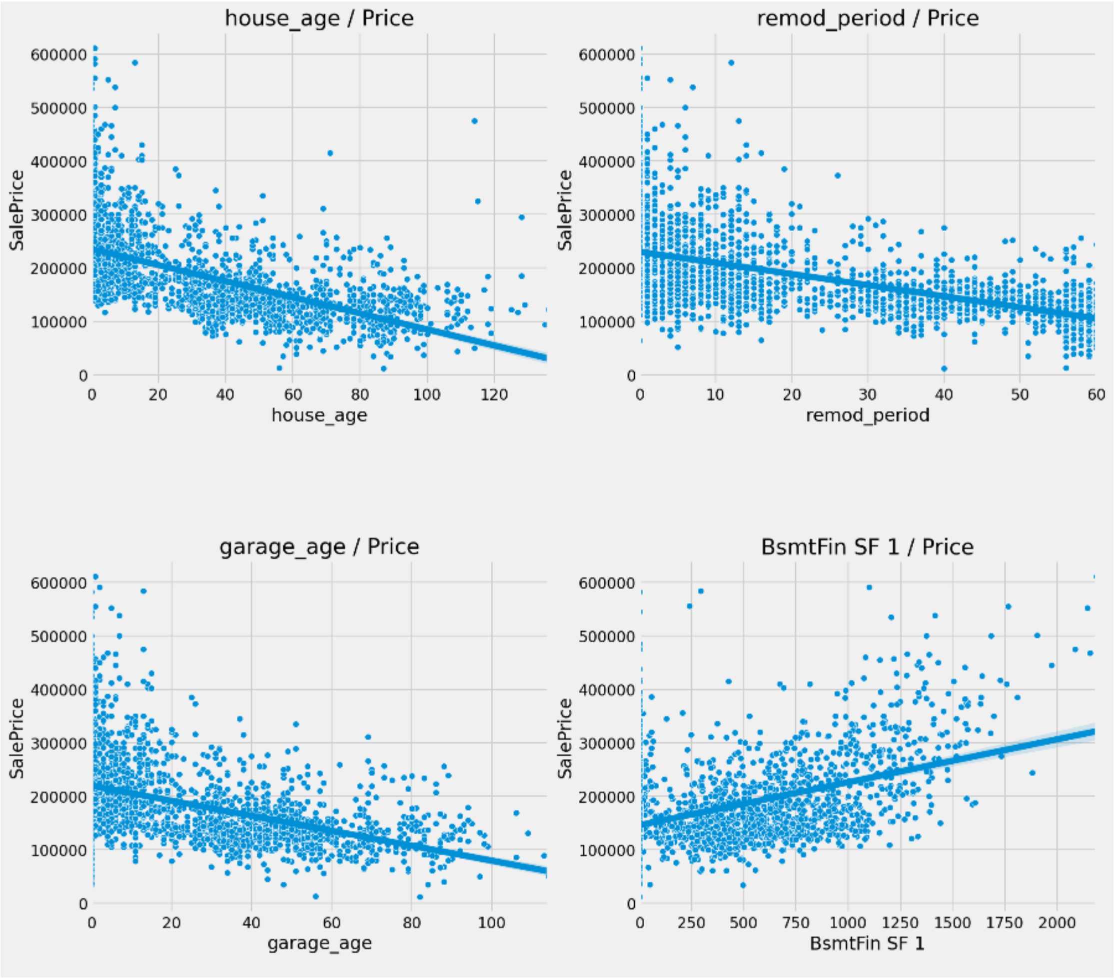
Features are filtered out if its correlation rate with Sale Price is >=0.4.



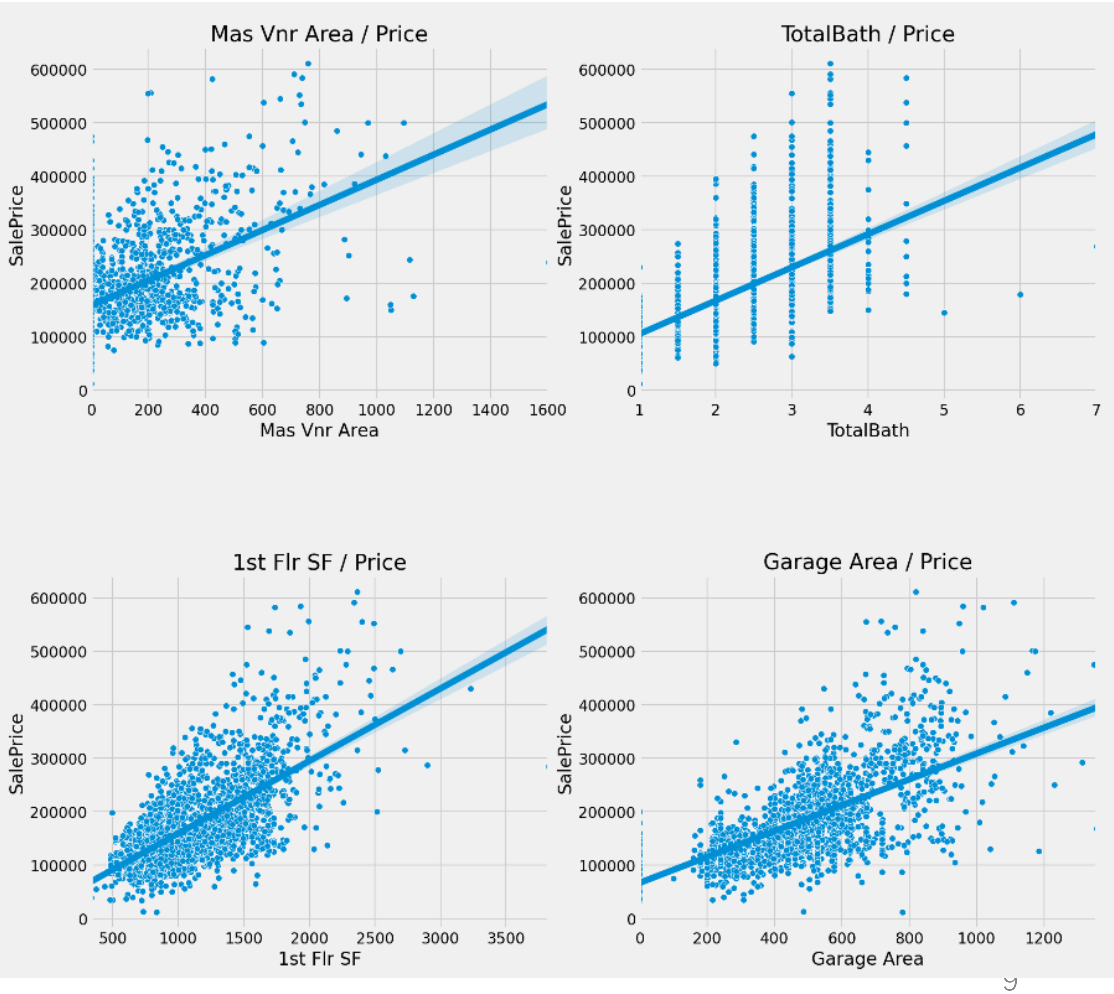
Data Analysis Processes

- **Features Selection - visualization for selected features**

Continuous data with scatter plot



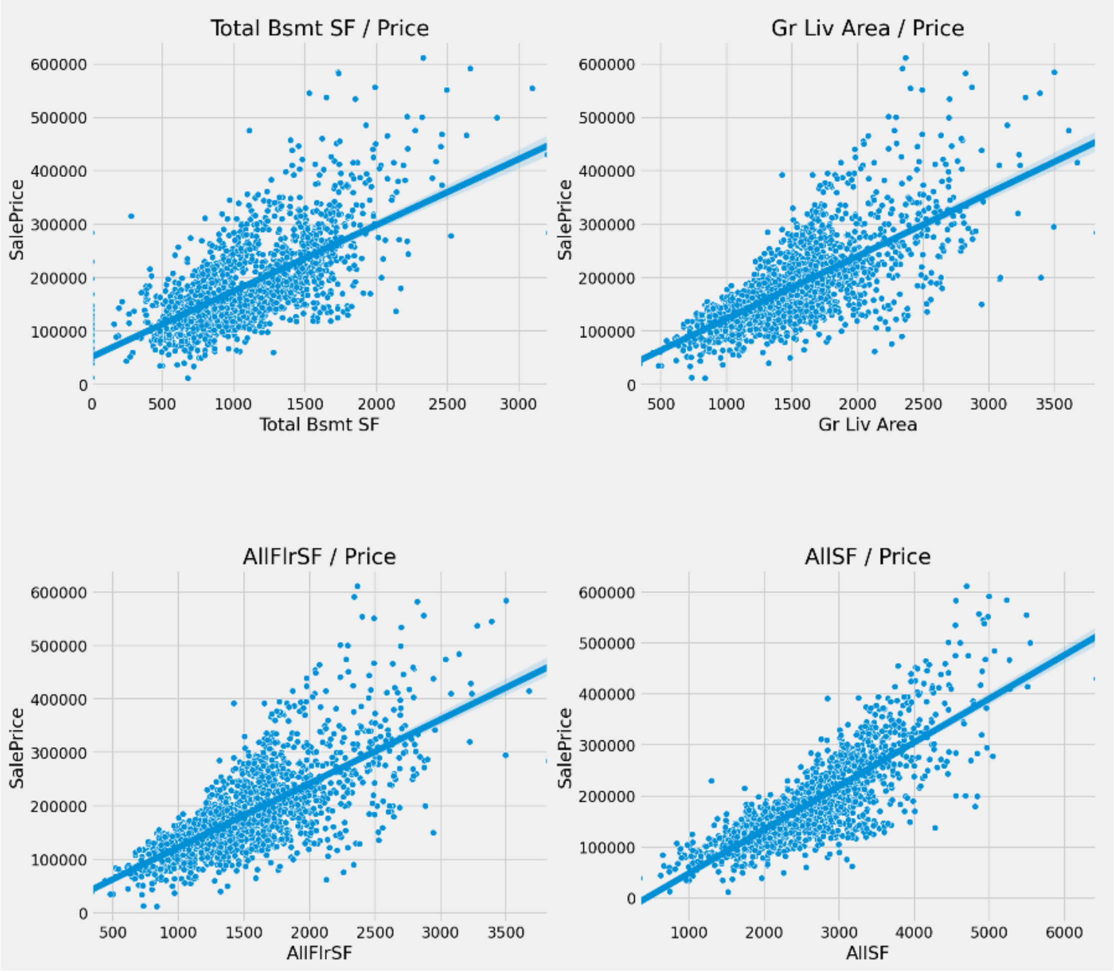
Continuous data with scatter plot



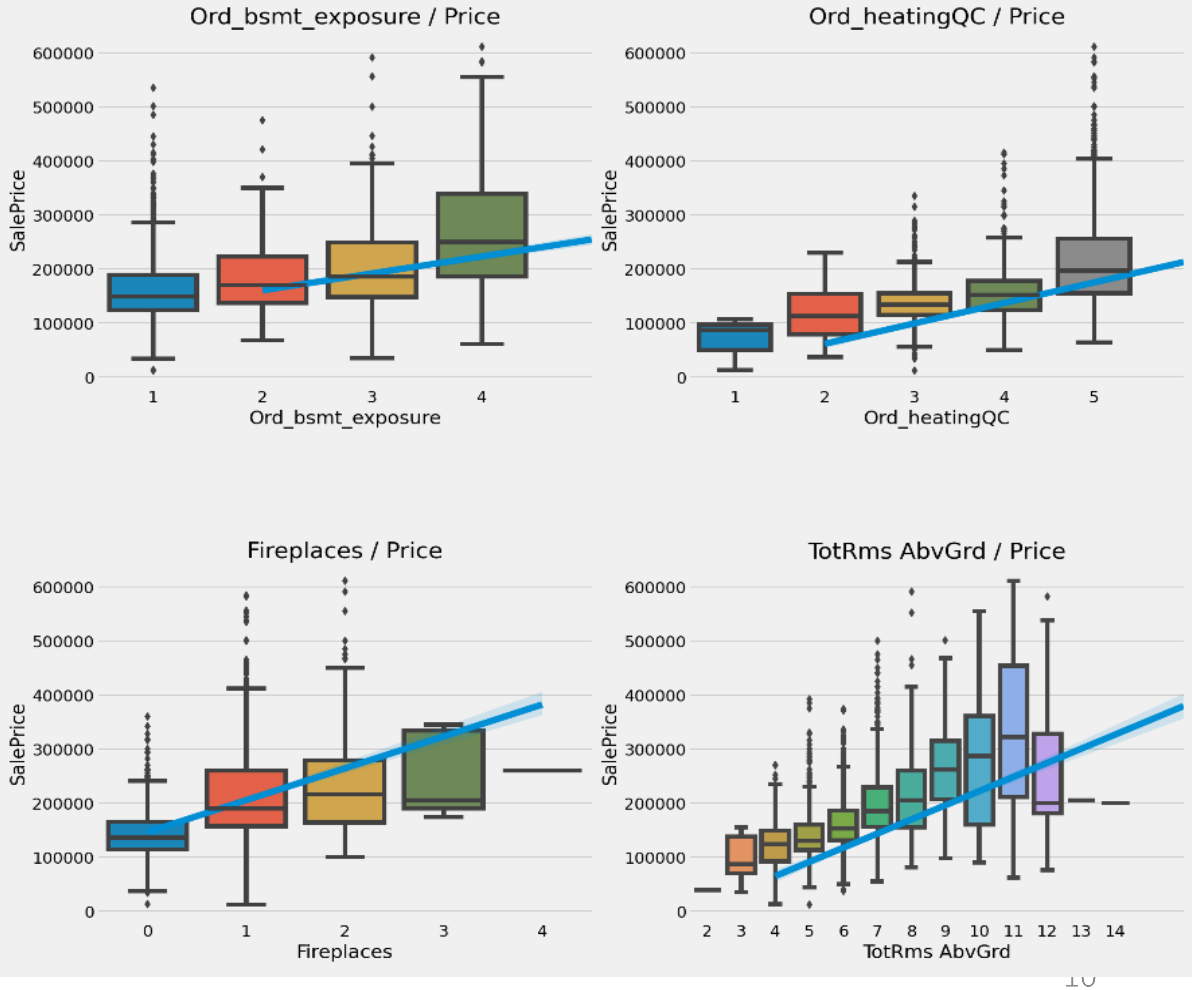
Data Analysis Processes

- **Features Selection - visualization for selected features**

Continuous data with scatter plot



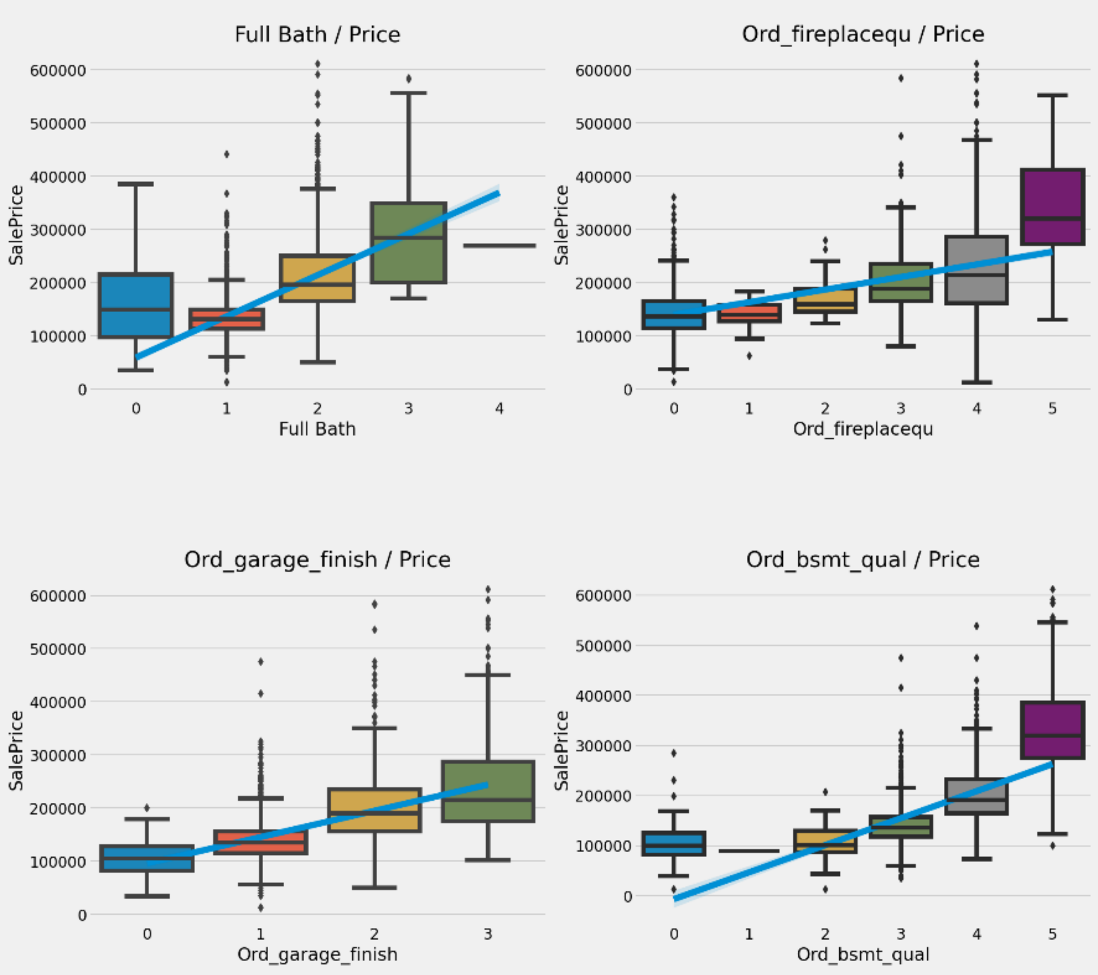
Discrete data with box plot



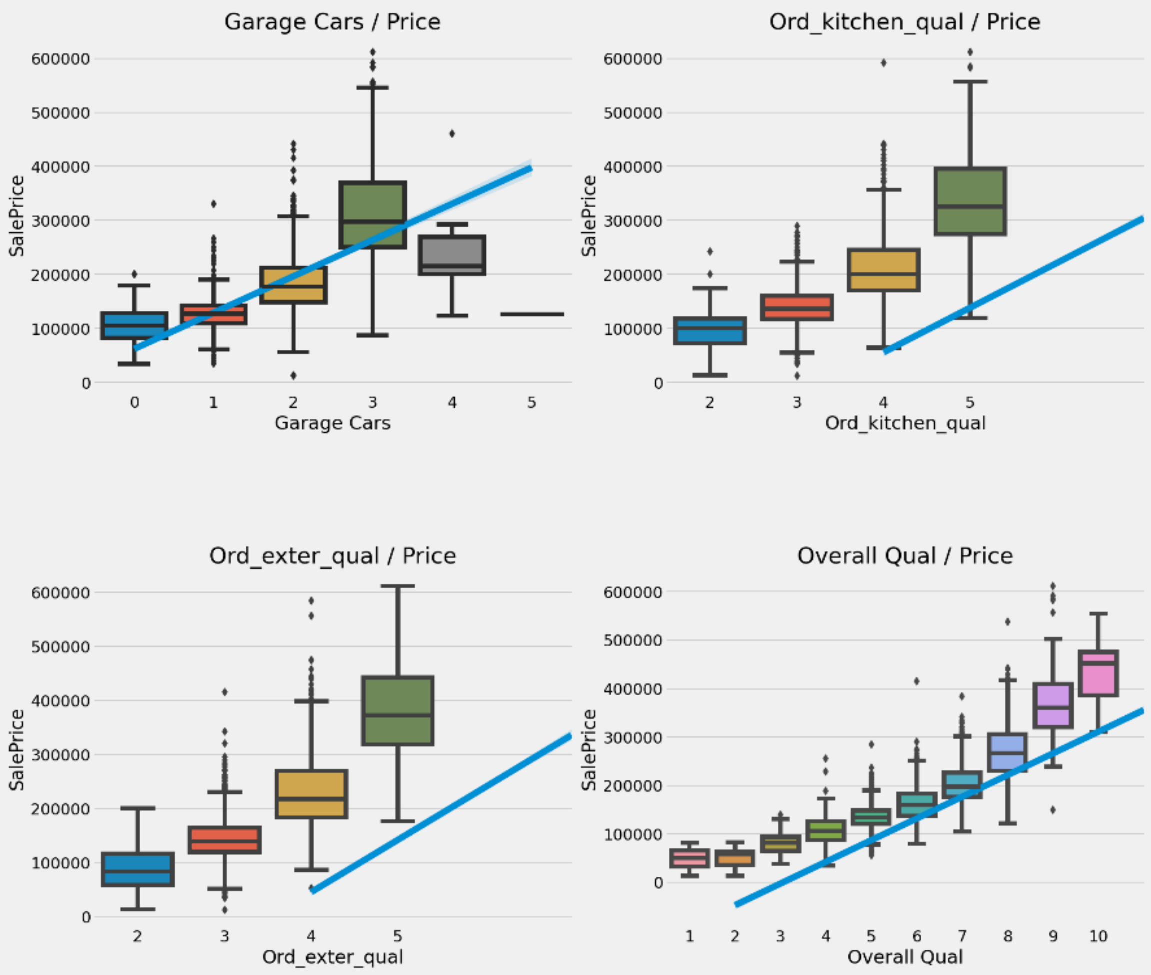
Data Analysis Processes

- **Features Selection - visualization for selected features**

Discrete data with box plot



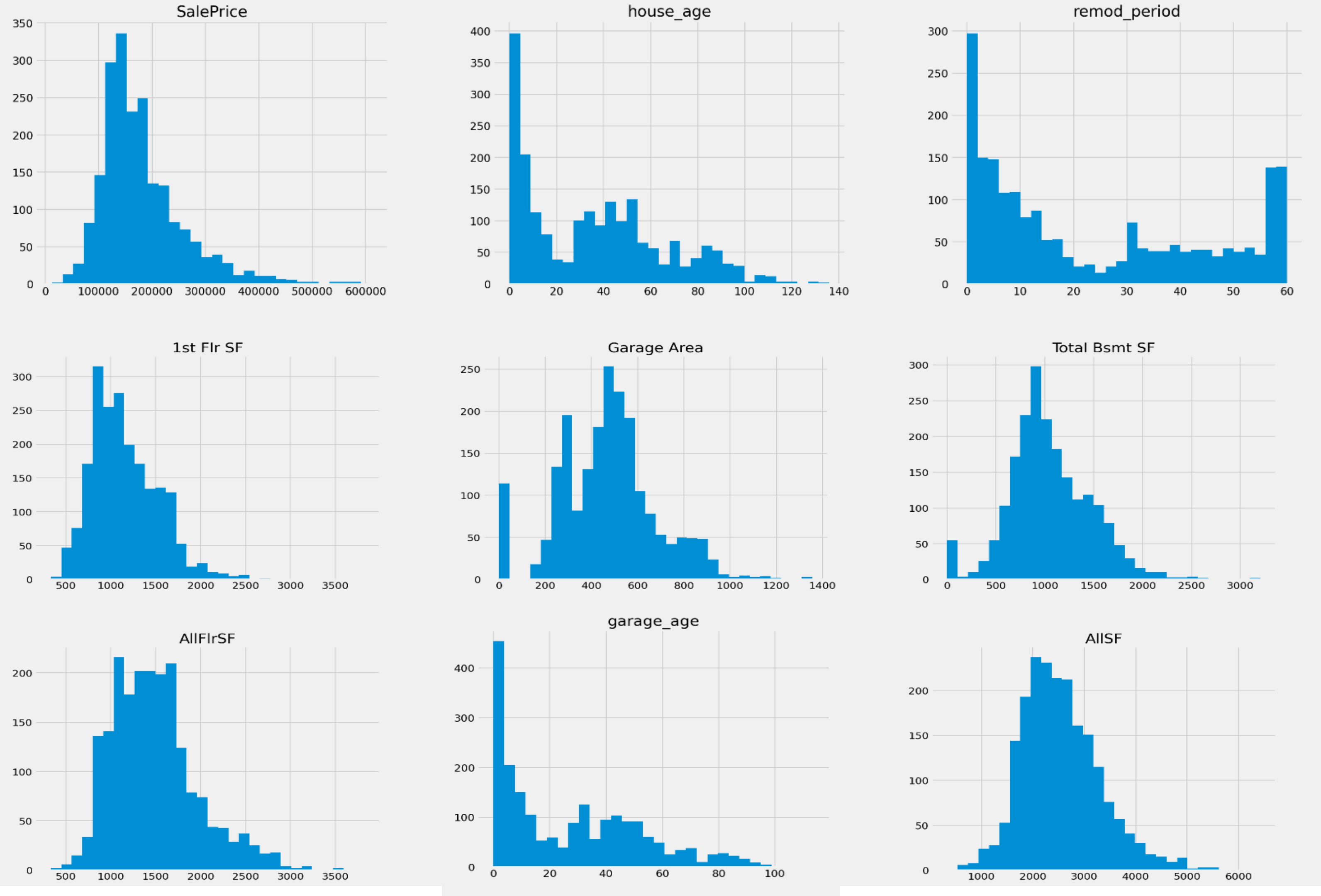
Discrete data with box plot



Data Analysis Processes

- **Features Selection - visualization for selected features**

Histogram plots – non-normal distribution

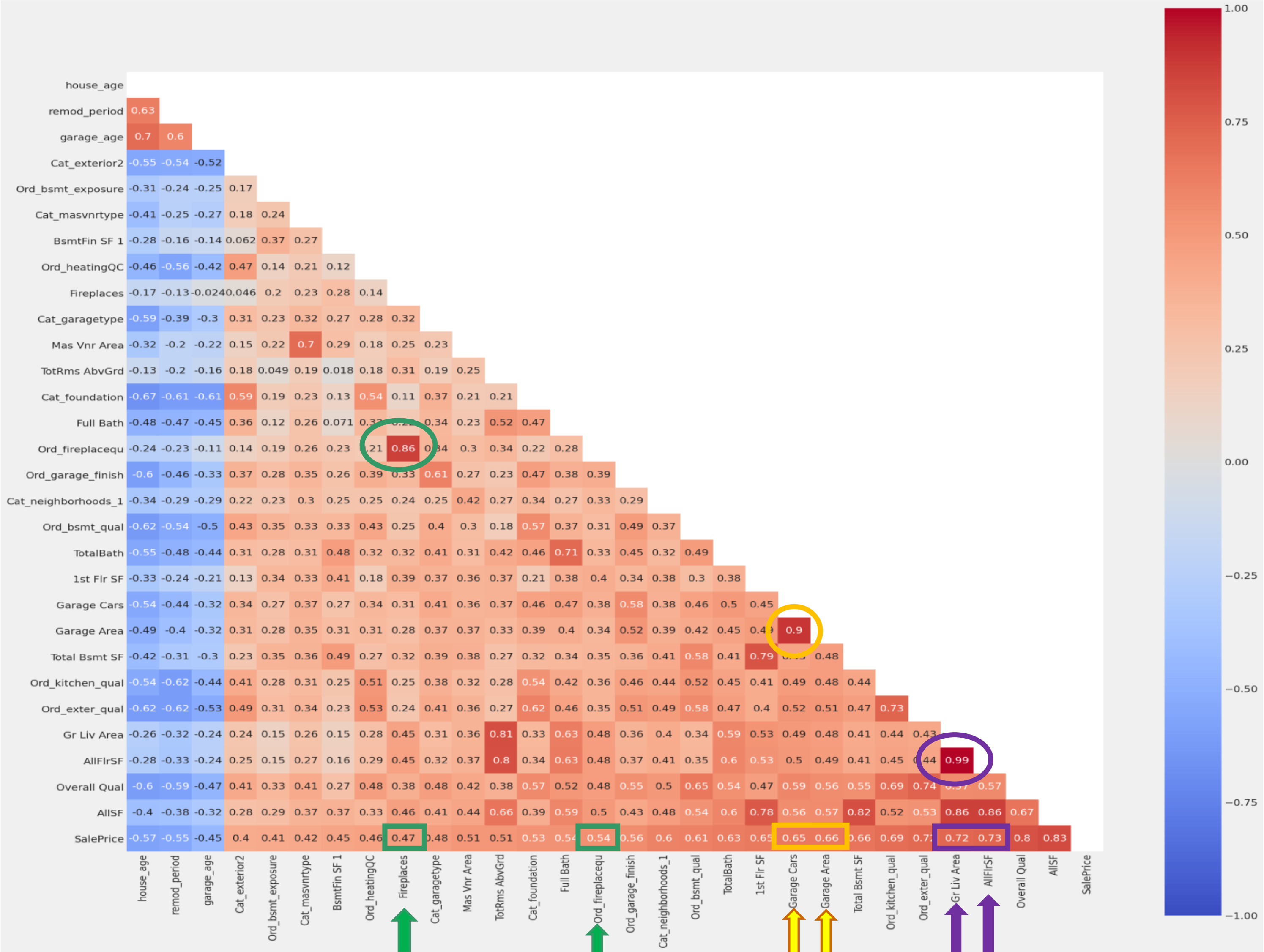


Data Analysis Processes

- **Features Selection - Check collinearity within features**

High collinearity pairs

1. Gr Liv Area vs. AllFltSF
   - ➢ Drop Gr Liv Area
2. Garage cars vs. Garage Area
   - ➢ Drop Garage Cars
3. Firepalces vs. ord_firepalcequ
   - ➢ Drop Firepalces



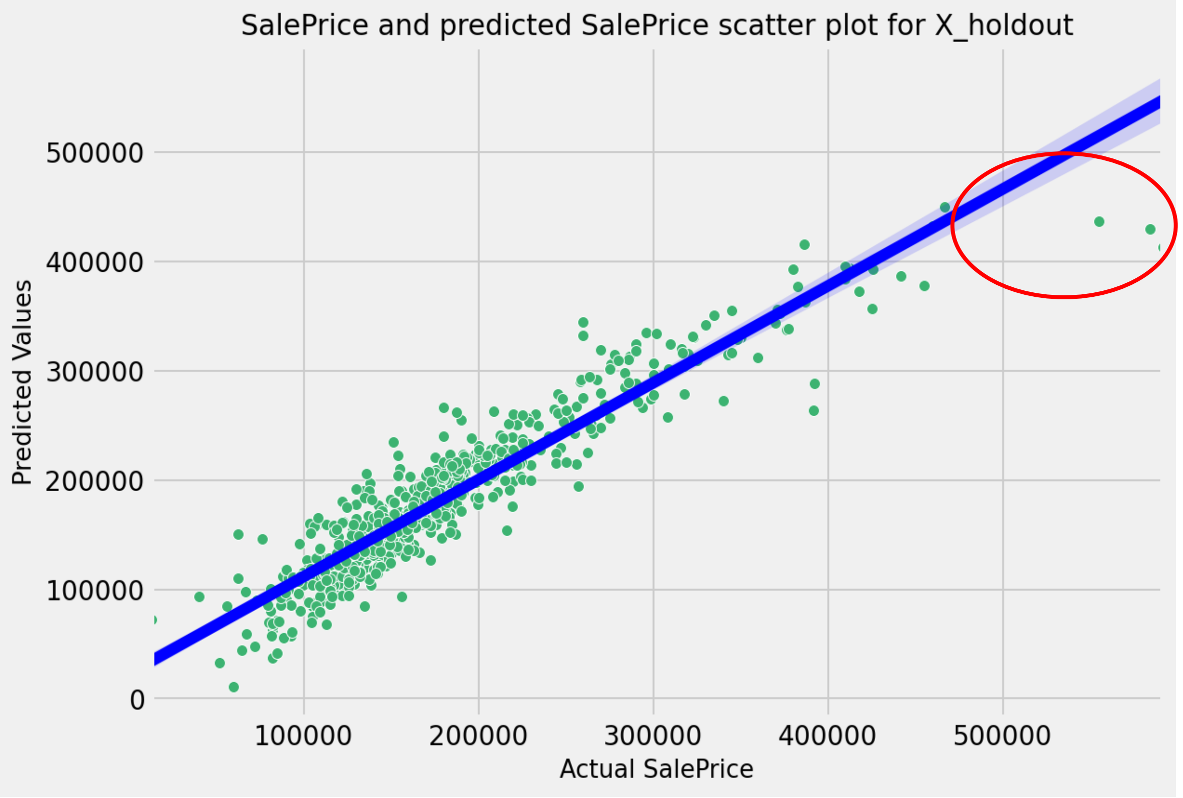
Data Analysis Processes

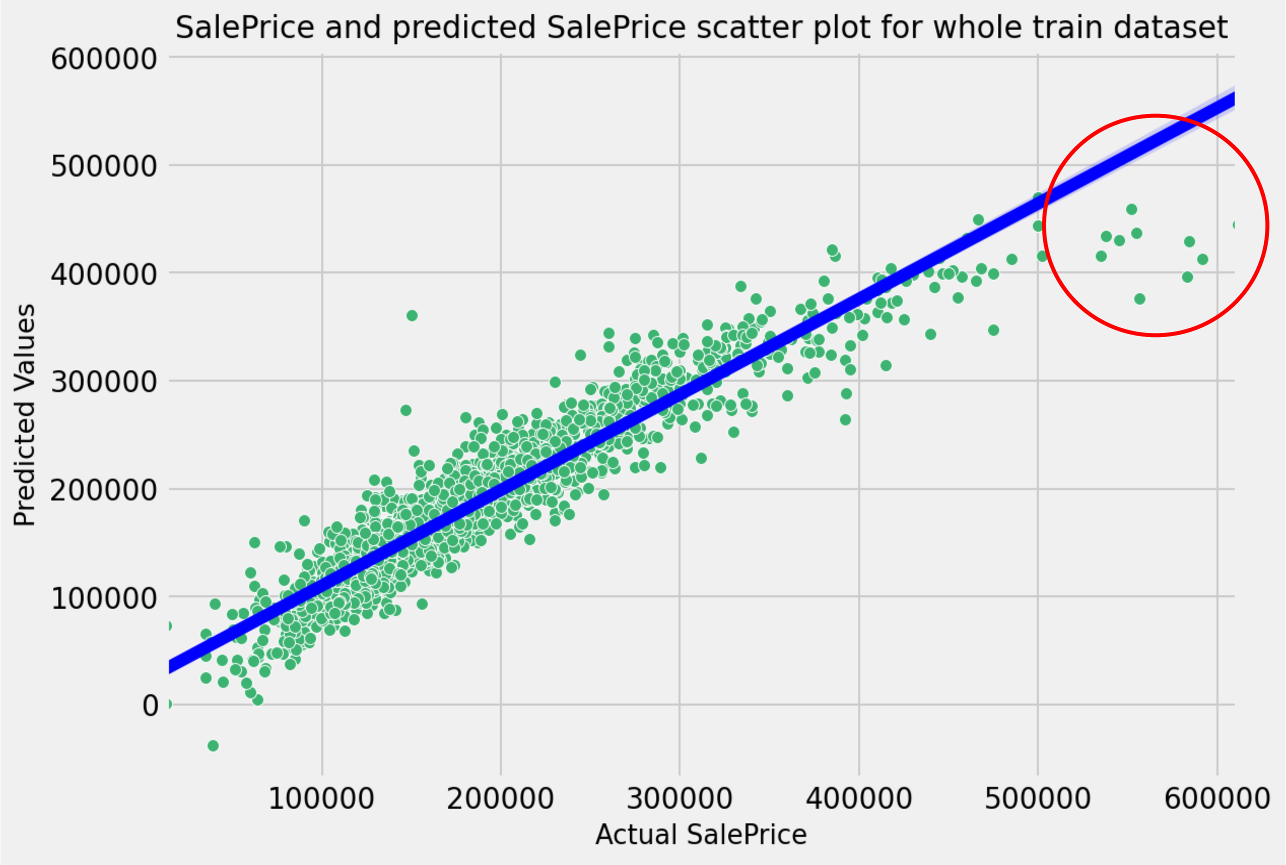
- **Model Verification - 1st round verification: Apply selected features**

Model	train RMSE	hold RMSE	train R2	hold R2
Linear Regression	27255.4138	28050.6253	0.8890	0.8752
Ridge Regression	27221.1833	27992.7968	0.8889	0.8753
Lasso Regression	**27207.6141**	**27975.5524**	0.8889	0.8749
ElasticNET Regression	**27207.6141**	**27975.5524**	0.8889	0.8749

Data Analysis Processes

- **Model Verification - 1st round verification: Apply selected features**
  - Plots with best model in 1st round verification, the model fit well for SalePrice from 0 to 500000, but not fit well in higher SalePrice which tends to overestimate.
  - Add power 2 (square) features to verify in 2nd round.





Data Analysis Processes

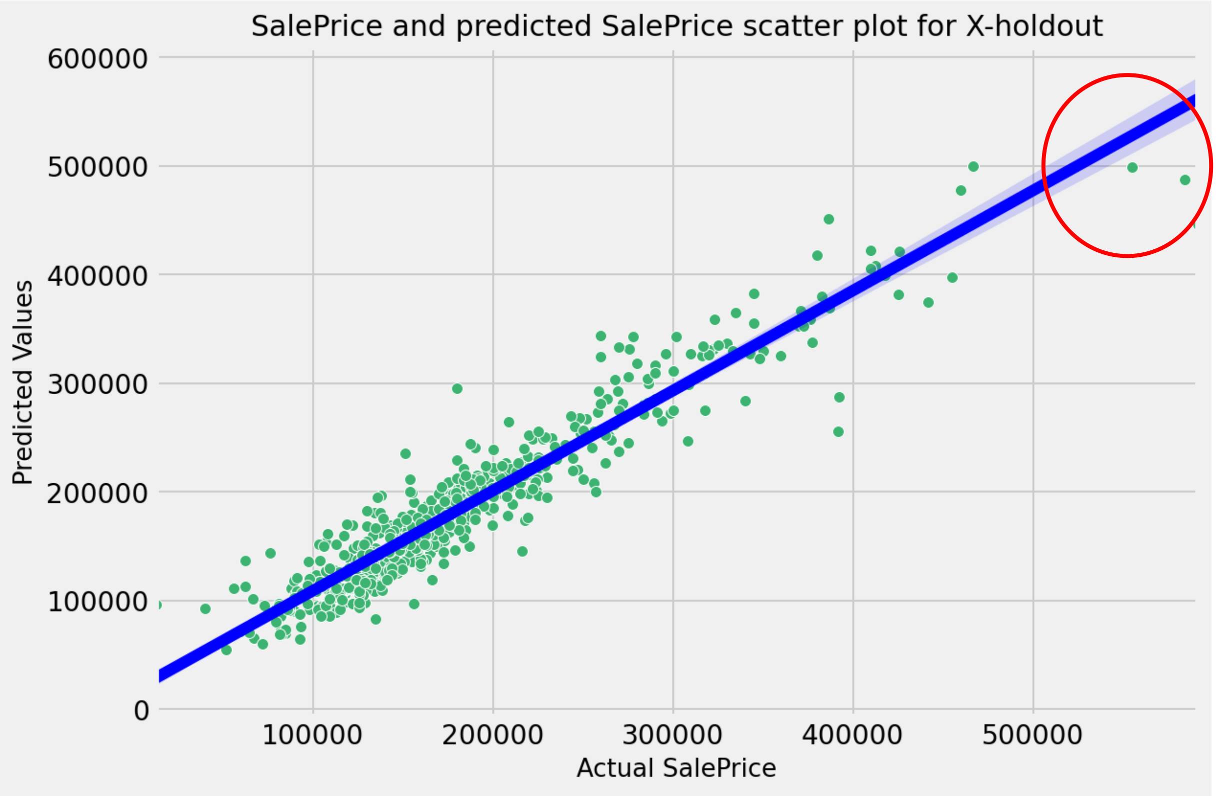
- **Model Verification – 2nd round verification: Apply Add power 2 (square) features**

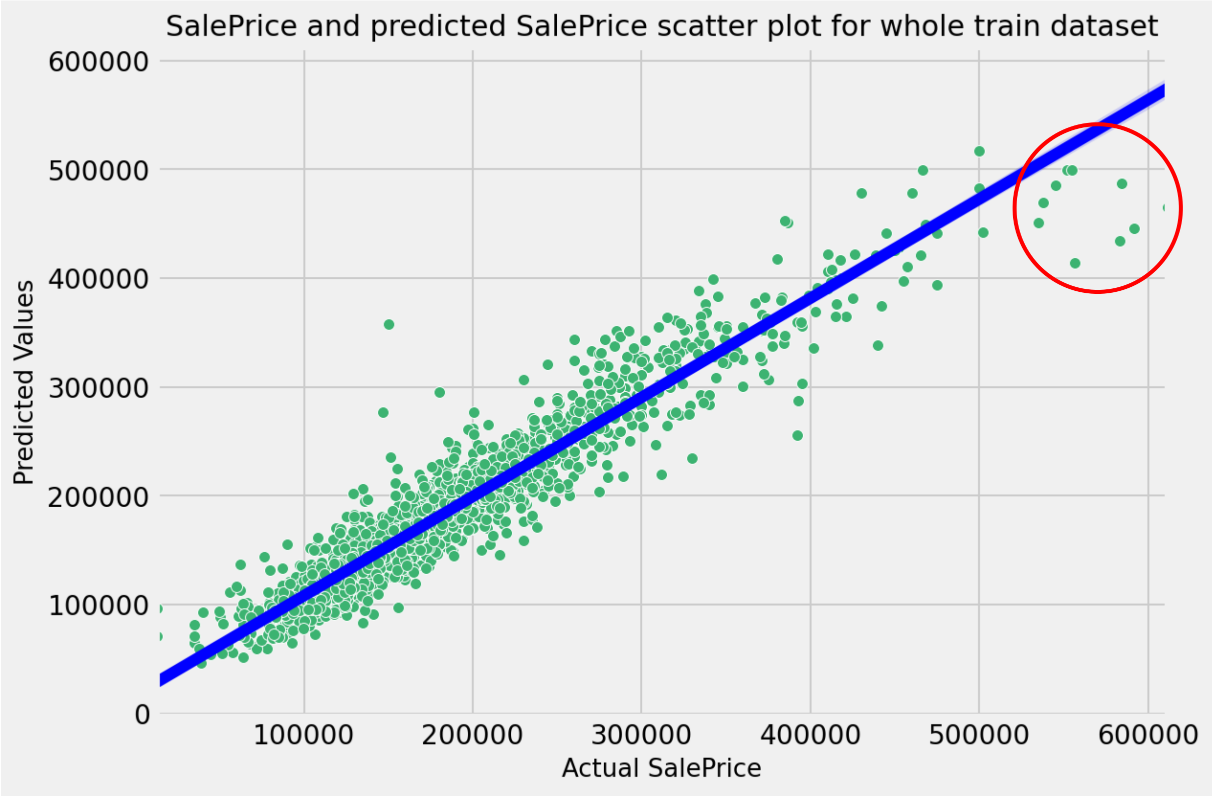
Model	train RMSE	hold RMSE	train R2	hold R2
Linear Regression	24795.9652	25637.2296	0.9103	0.8990
Ridge Regression	**24761.1563**	**25348.8987**	0.9101	0.8990
Lasso Regression	24809.6161	25616.3006	0.9097	0.8980
ElasticNET Regression	24809.6161	25616.3006	0.9097	0.8980

- Observe significant reduce for RMSE from **27207 to 24761.**

Data Analysis Processes

- **Model Verification - 2nd round verification: Apply Add power 2 features**
  - Plots with best model in 2nd round verification shows much improvement fit for higher SalePrice but still not fit well.
  - Add higher power (i.e. 3) to verify whether further improvement.





Data Analysis Processes

- **Model Verification – 3rd round verification: Add power 3 features**

Model	train RMSE	hold RMSE	train & hold RMSE diff	train R2	hold R2
Linear Regression	24687.4268	26424.0999	-1736.6731	0.9135	0.9028
Ridge Regression	**24693.2033**	**25030.6911**	-337.4878	0.9111	0.9010
Lasso Regression	**24651.7675**	**25326.342**	-674.5745	0.9111	0.9006
ElasticNET Regression	**24651.7675**	**25326.342**	-674.5745	0.9111	0.9006

- Observe RMSE score improved but not that much.
- Higher power features caused high variance in linear regression.
- Can not choose best model due to **good train score** for **Lasso/ElasticNET regression but good hold score for ridge regression.**

Data Analysis Processes

- **Model Verification – 3rd round verification: Add power 3 features**
  - Many features with Zero coeffects in Lasso regression
  - May due to high collinearity between features and their power 2/power 3

	Coefficient
Overall Qual	0.0
AllSF	0.0
Ord_exter_qual_s2	-0.0
Overall Qual_s2	-0.0
AllSF_s2	0.0
Ord_kitchen_qual_s2	-0.0
AllFlrSF_s3	0.0
Garage_Area_s3	0.0
Cat_exterior2	-0.0

Data Analysis Processes

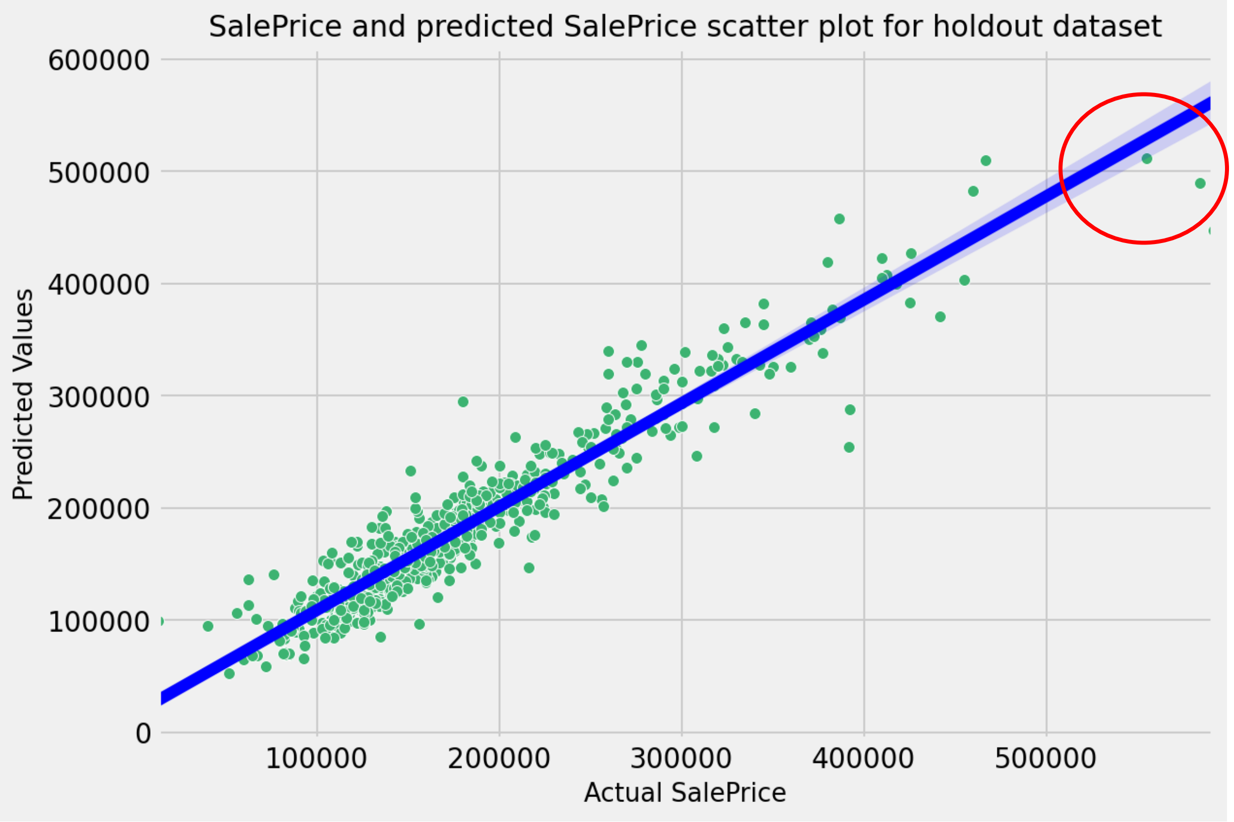
- **Model Verification – 4th round verification: Drop zero coeffects from 3rd round**

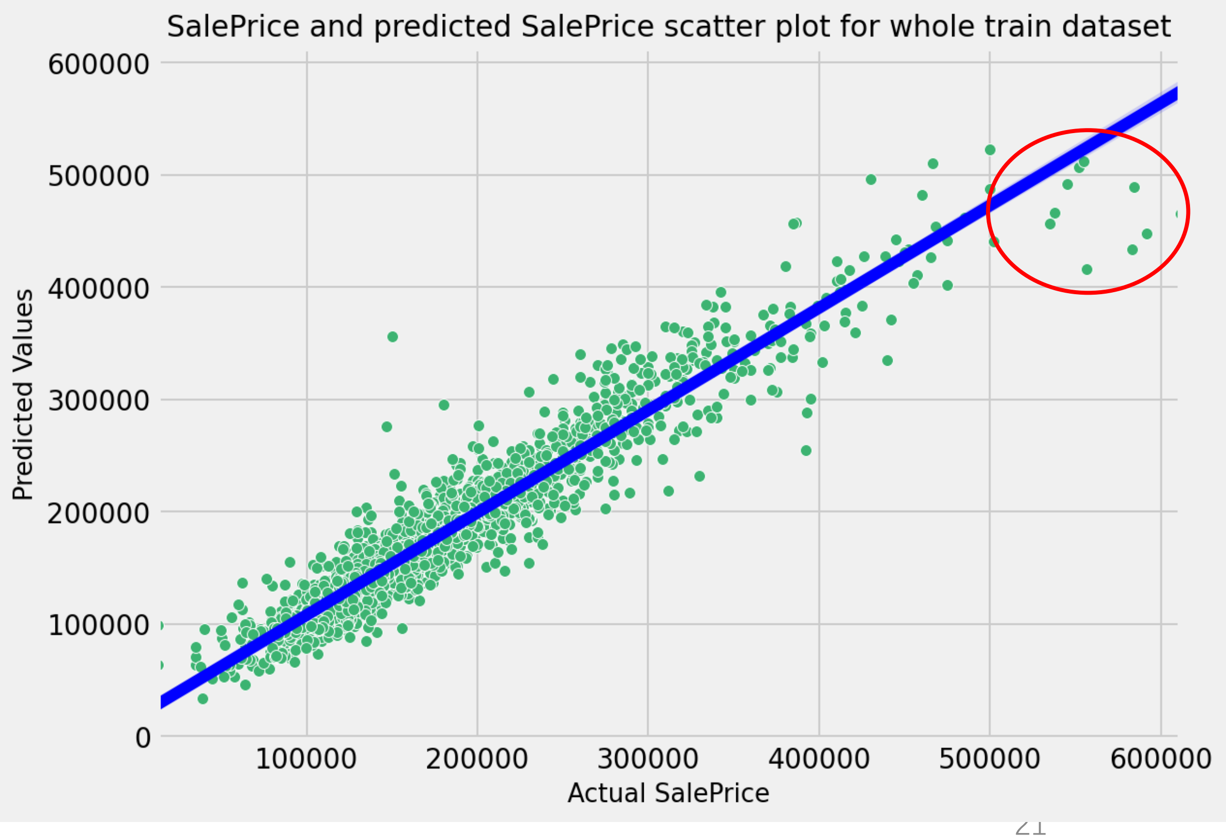
Model	train RMSE	hold RMSE	train & hold RMSE diff	train R2	hold R2
Linear Regression	24593.4239	25173.9070	-580.4831	0.9115	0.9010
Ridge Regression	**24546.7903**	**24833.6543**	**-286.864**	0.9112	0.9011
Lasso Regression	24600.6990	25063.6965	-462.9975	0.9114	0.9009
ElasticNET Regression	24600.6990	25063.6965	-462.9975	0.9114	0.9009

- Observe the reduced gap between train data RMSE score and holdout data RMSE score, especially Ridge regression.
- Choose **Ridge regression as best model** which can fit well for both train and hold data.
- **For whole train dataset, RMSE for best mode is 24360.1724 and R2 is 0.9057.**

Data Analysis Processes

- **Model Verification - 4th round verification: Drop zero coeffects from 3rd round**
  - Plots for higher SalePrice is similar to plots in 2nd round verification.
  - Higher power features can still improve the model, but may cause high variance and collinearity issue.

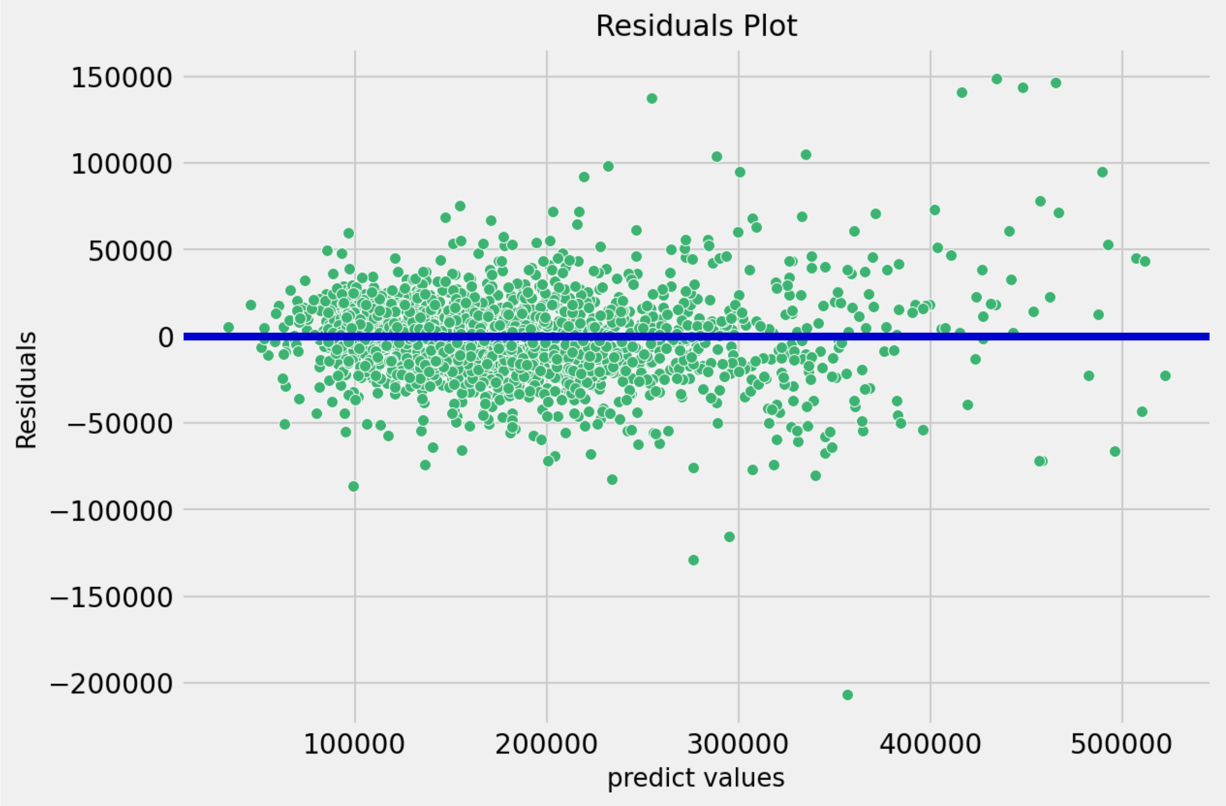


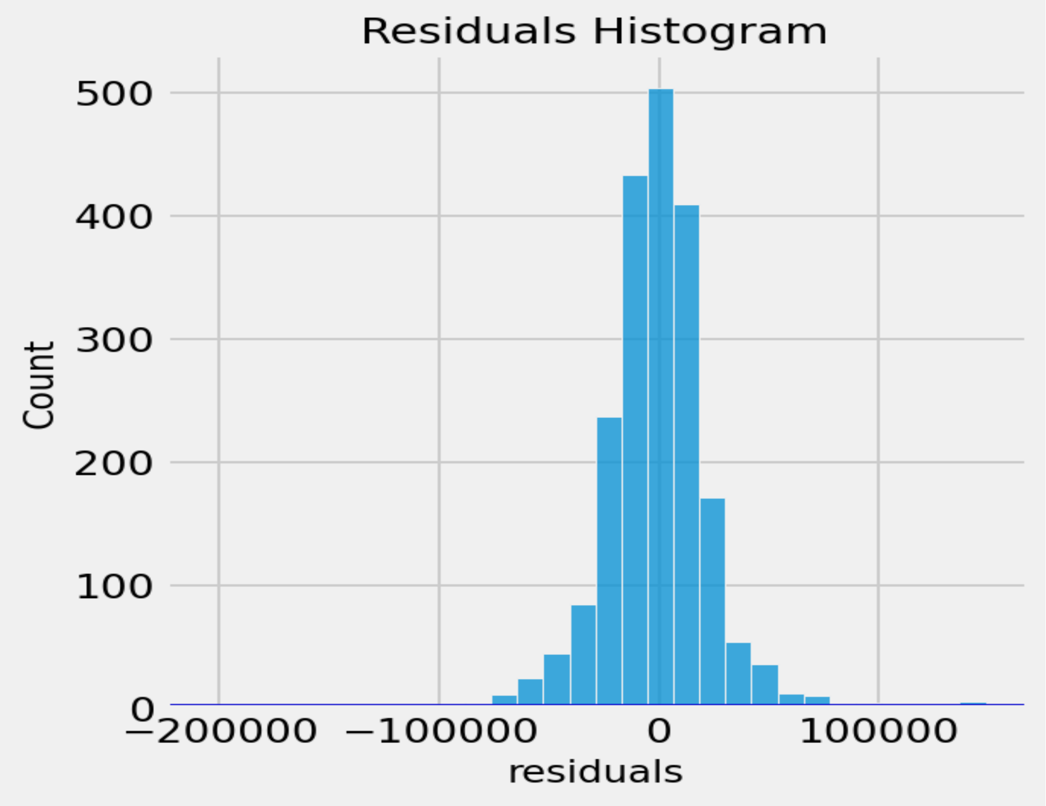


Data Analysis Processes

- **Model Verification - Residual plot with best model**

  - **Residuals for whole train data set scatters around zero and nearly normally distribution.**





Conclusion and Recommendations

- The model created performs well for 90.57% of the variation in Sale Price
- It does not fit well for extreme high SalePrice.
- Power 2/3 features can help to improve the prediction, but higher power features may raise high variance and collinearity between features.

- From this model, we can make some recommendations for homeowners to increase their property value.
  1. Maintain overall house quality including kitchen, internal and external of house etc.
  2. Increase floor area if possible
  3. Make house well-renovated as good living quarters including basement area
  4. With builtIn or attached garage
  5. New houses and newly-renovated houses are more valuable.
  6. The houses in neighborhoods, such as neighborhoods Stone Brook, Northridge Heights, Veenker, Northridge, Green Hills are more valuable.