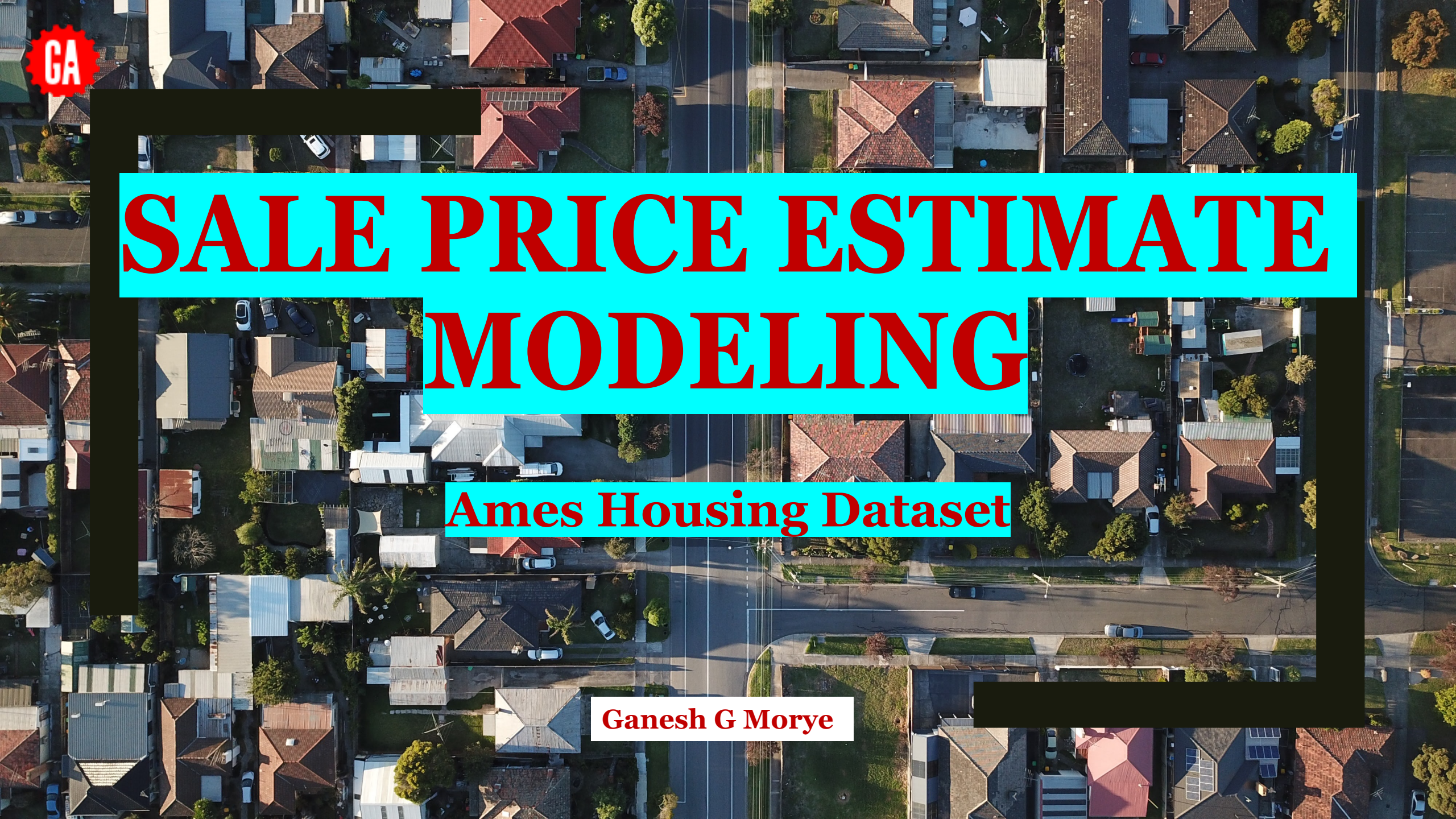


SALE PRICE ESTIMATE MODELING

Ames Housing Dataset

Ganesh G Morye

Contents

Problem Statement

Introduction

Datasets Overview

Model Building Workflow

Conclusions

Build a Multiple Linear Regression model to estimate sale price for a house listed in the city of Ames, Iowa

Problem Statement

Introduction

- Buyers, sellers and investors of real estate properties are always motivated to strike the best “deal”
- Before closing the deal, parties involved need to feel confident about their decision
- Tools like Zestimate are available to provide a certain level of comfort to the involved parties in their decision
- A house is a physical structure with quantifiable features such number of bedrooms, baths, kitchens, pool, and fireplaces etc.
- Linear regression models are one of the machine learning techniques to model a sale price of a property

Datasets Overview

- Datasets provided
  - *train.csv – 2051 properties data with 80 features and sale price*
  - *test.csv – 878 properties data with 80 features and no sale price*
  - *Listed properties were sold between the years 2006-2010*
- *train.csv* is used to train and fit the model
- *test.csv* is used to make predictions and submit them to Kaggle to score the model
- Models are scored using the root-mean-squared-error(RMSE) metric
- Model choices are limited to Multiple linear regression and their regularization based on Lasso, Ridge and ElasticNet

Model Building Workflow

- Understanding the data - Data dictionary
- Data Cleaning
- Exploratory Data Analysis– Data Visualization
- Feature Selection and Engineering
- Model Evaluation

Data Dictionary

Numerical Variables

- *Discrete*
- *Continuous*

Categorical Variables

- *Nominal*
- *Ordinal*

Dean De Cock has provided a detailed description of all the features associated with a particular listing.

Data Cleaning

- 10,000 missing values in the *train.csv* dataset of 160,000 values.
- At least one null value in every row
- Top 5 features with the most missing value counts are:
- 27 features with at least one missing value
- Data cleaning involved imputing these missing values

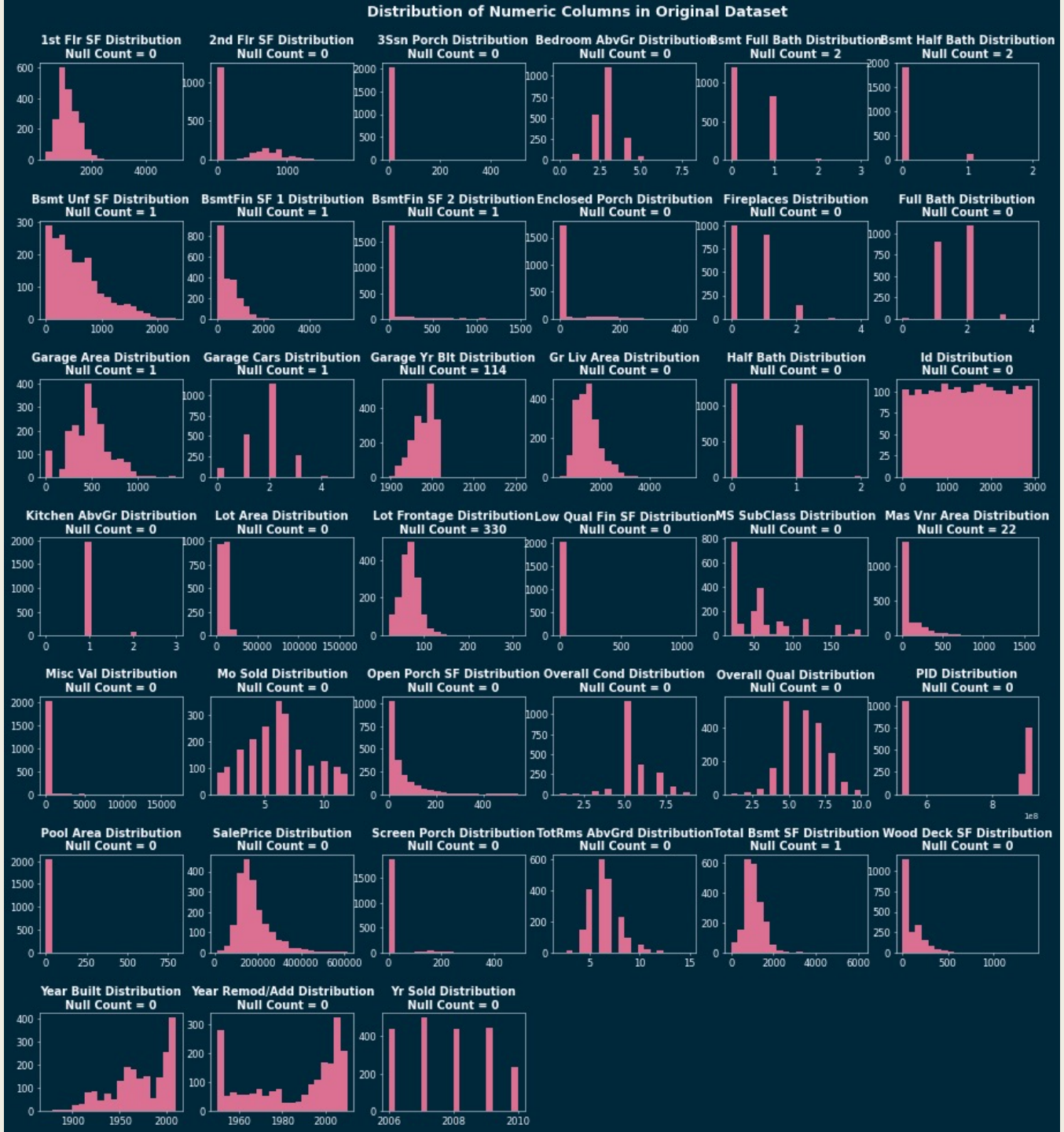
Feature	Count of Null Values
Pool QC	2042
Misc Feature	1986
Alley	1911
Fence	1651
Fireplace Qu	1000

Data Cleaning

- Steps involved in Data Cleaning
  - *Investigating the statistics*
  - *Generating Visualizations to look at the distribution*
  - *Inspecting related features to derive possible values*
  - *Functions to streamline the handling of both the test and train datasets in a consistent manner*
- E.g.
  - *7 features related to the Garage*
  - *Out of the 7, 6 of them have 114 rows of missing values*
  - *These 114 rows have Garage Area information (mostly zero)*
  - *So, the missing garage related features are just an indication that the property has no garage*
  - *Missing values are imputed using this association*

Data Visualizations

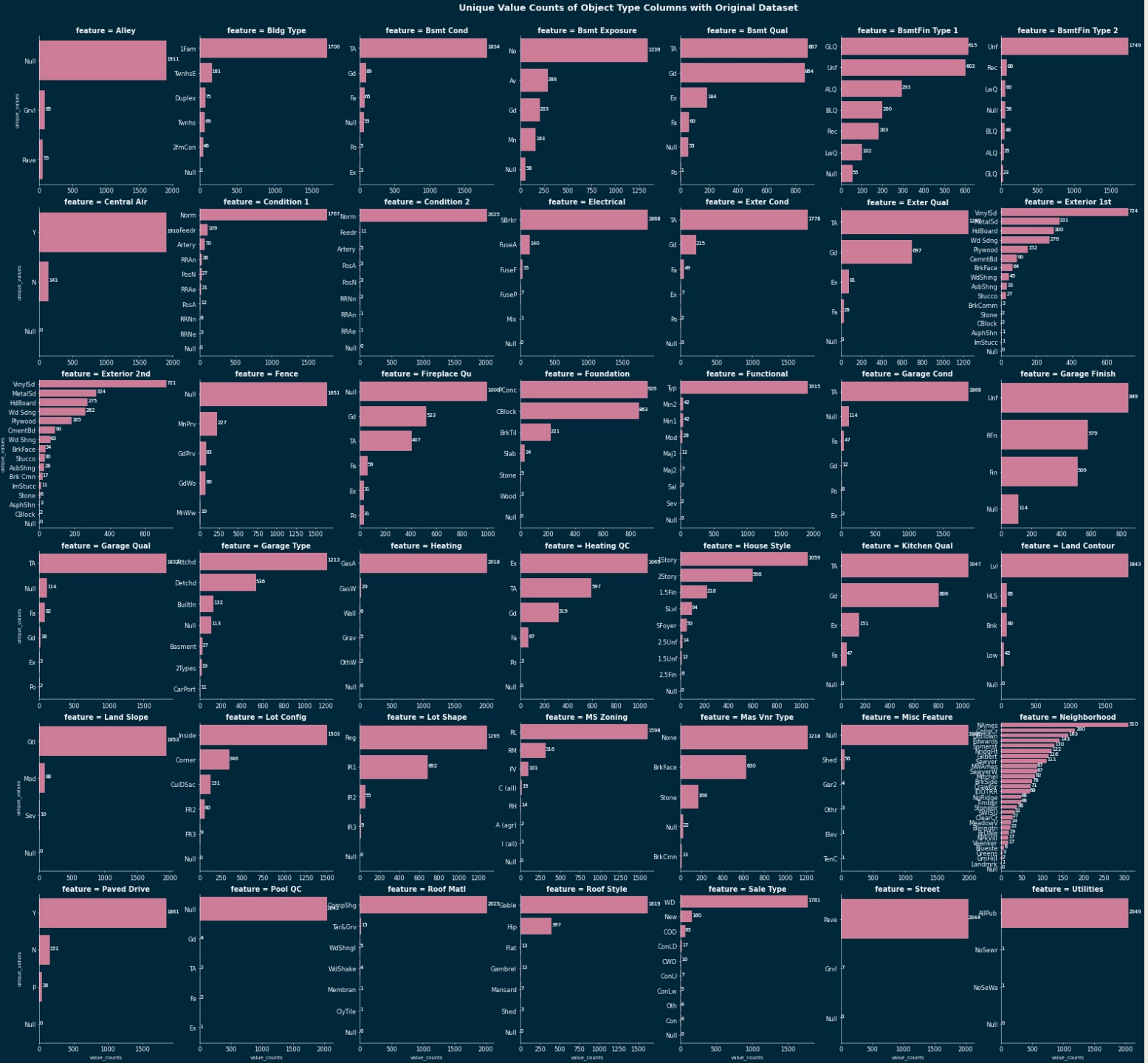
- The relationship between predictors and the response variables and amongst the predictors is investigated with aid of comprehensive suite of data visualizations
- Visualizations were grouped primarily based on their datatypes
- These visualizations aided in understanding
  - *Data quality spot-check*
  - *Distributions*
  - *Unique labels and counts*
  - *Strength and nature of the correlation to the sale price*
  - *Identifying group statistics such as IQR, outliers, mean, median*



Distribution of Numerical Features- Original Dataset *train.csv*

Model Building Workflow

Data Visualizations



Unique Value Counts in Object Features- Original Dataset *train.csv*

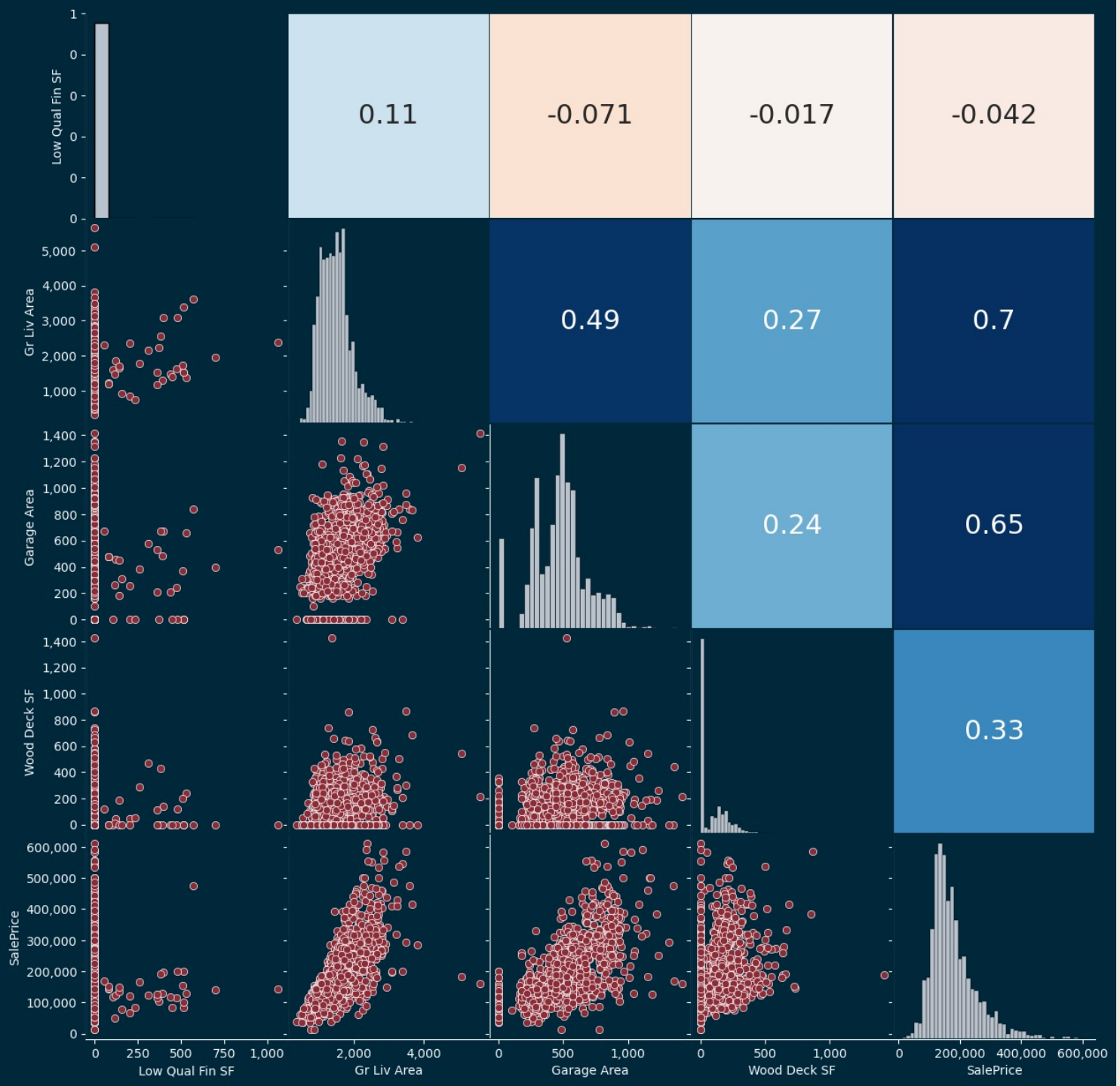
Model Building Workflow

Data Visualizations

Cleaned Dataset
Train vs Test
Numerical Features

Model Building Workflow

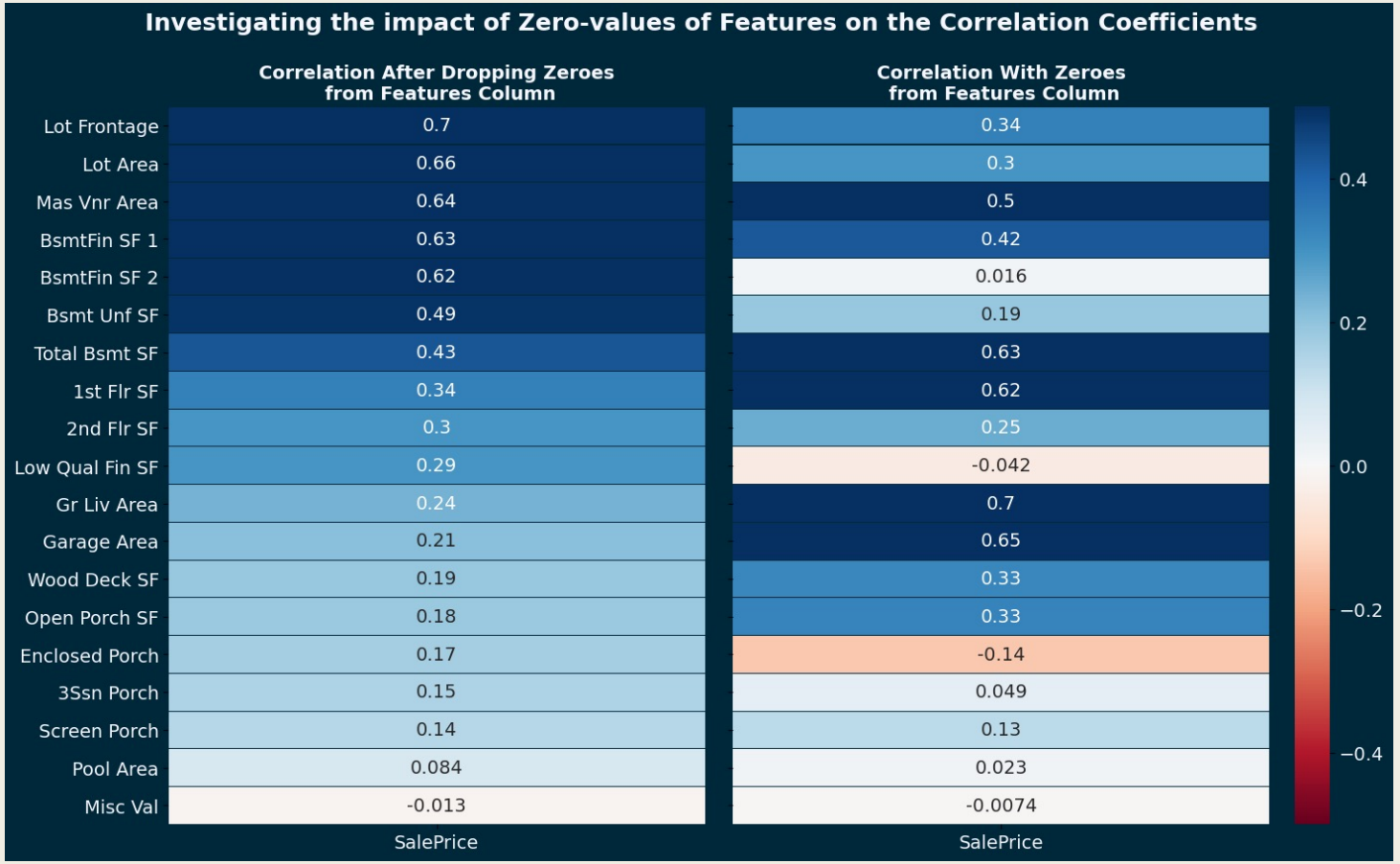
Data Visualizations



Correlation & Distributions Continuous Variables

Model Building Workflow

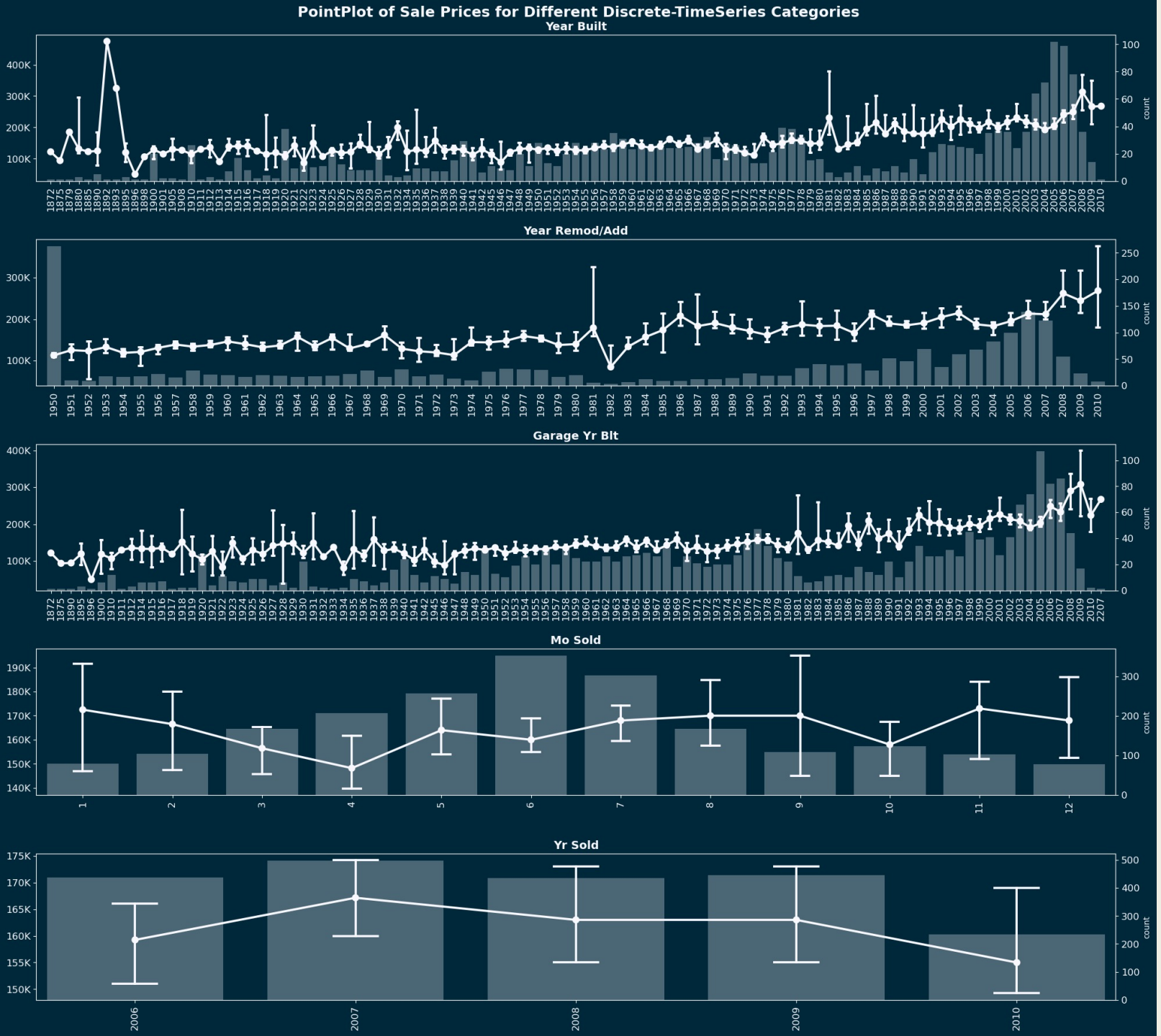
Data Visualizations



Correlation Matrix

Model Building Workflow

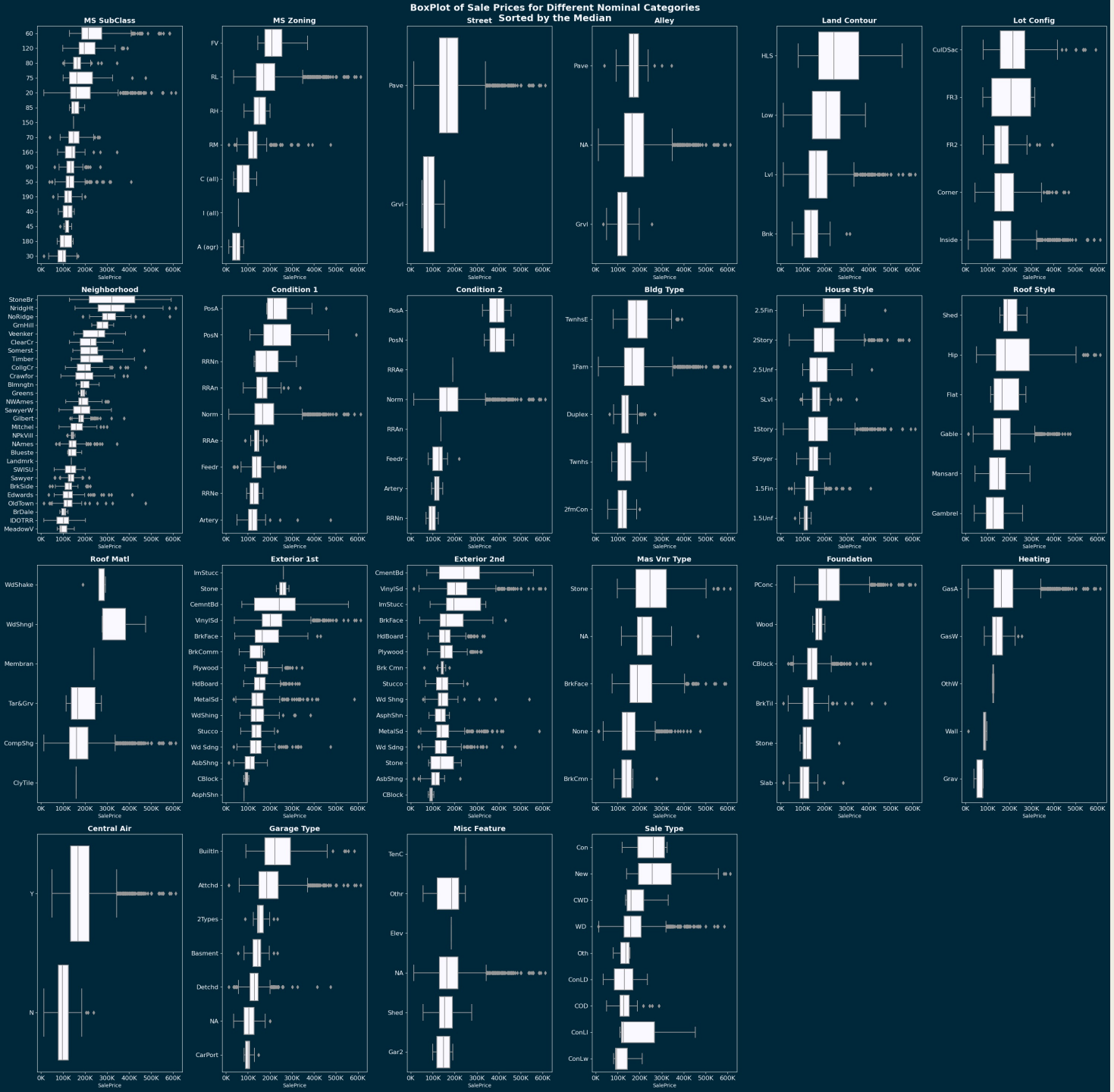
Data Visualizations



Point-Plot of Sale Prices for Time Series Variables

Model Building Workflow

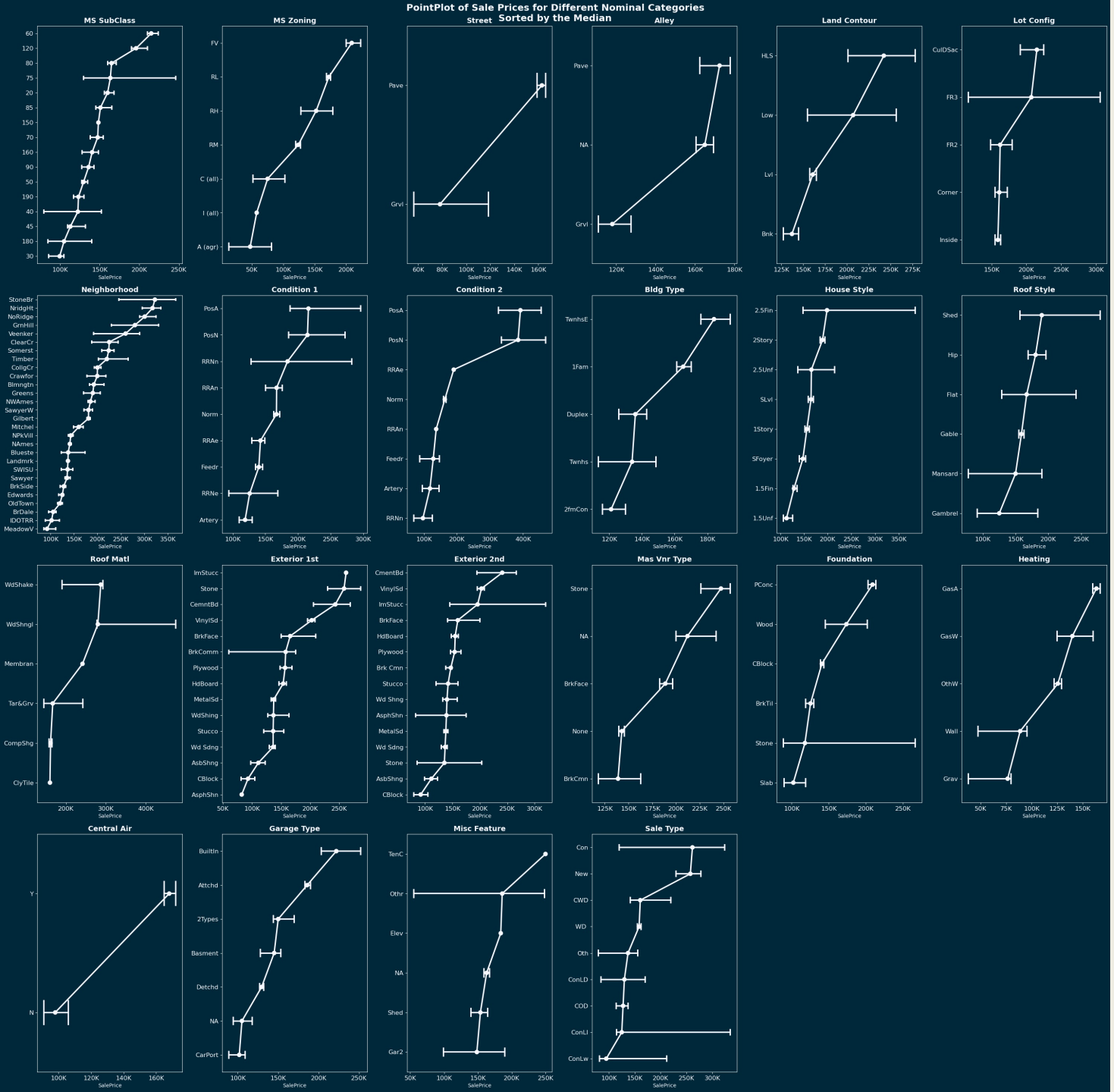
Data Visualizations



Box-Plot of Sale Price for Nominal Variables

Model Building Workflow

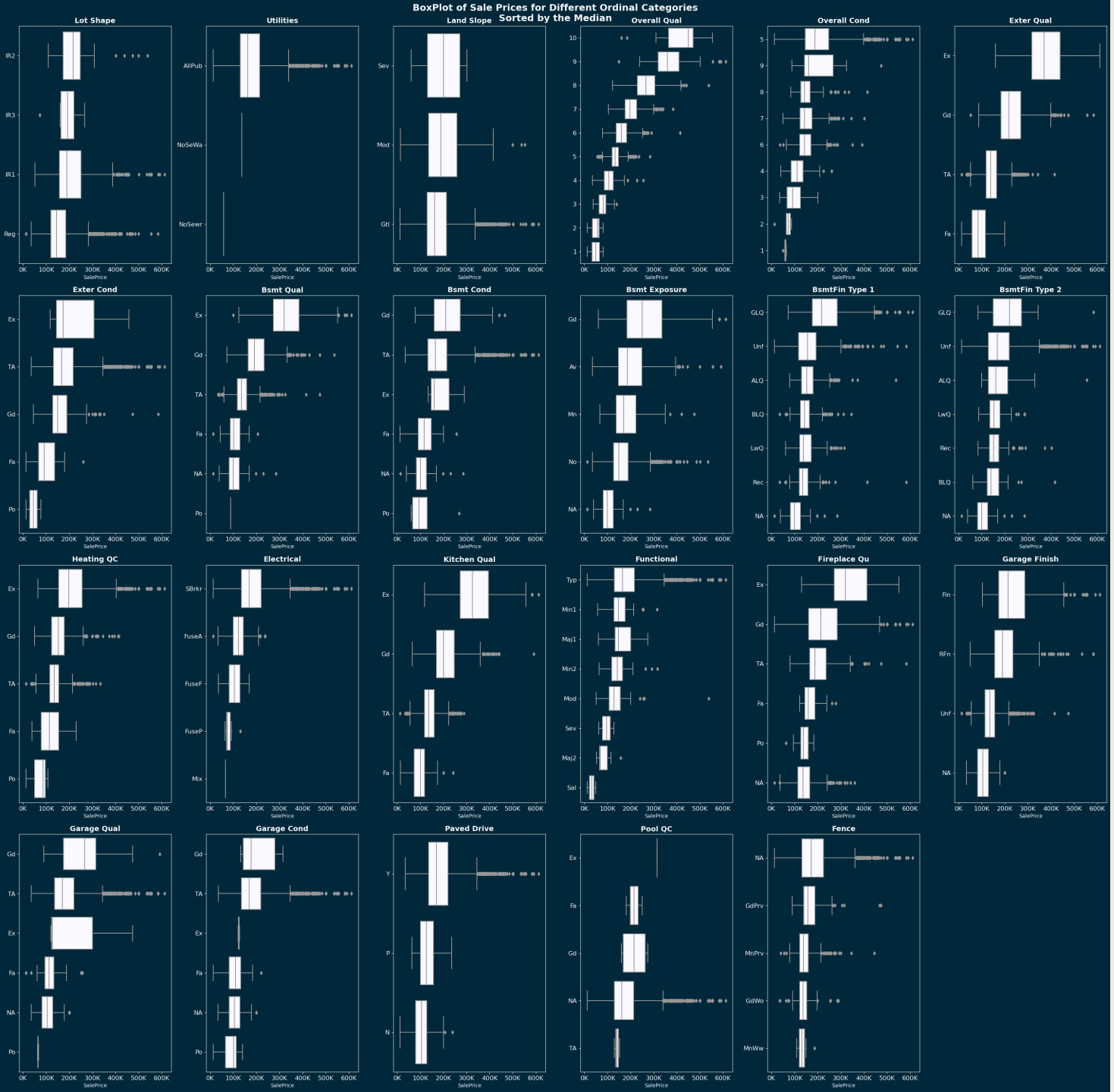
Data Visualizations



Point-Plot of Sale Price for Nominal Variables

Model Building Workflow

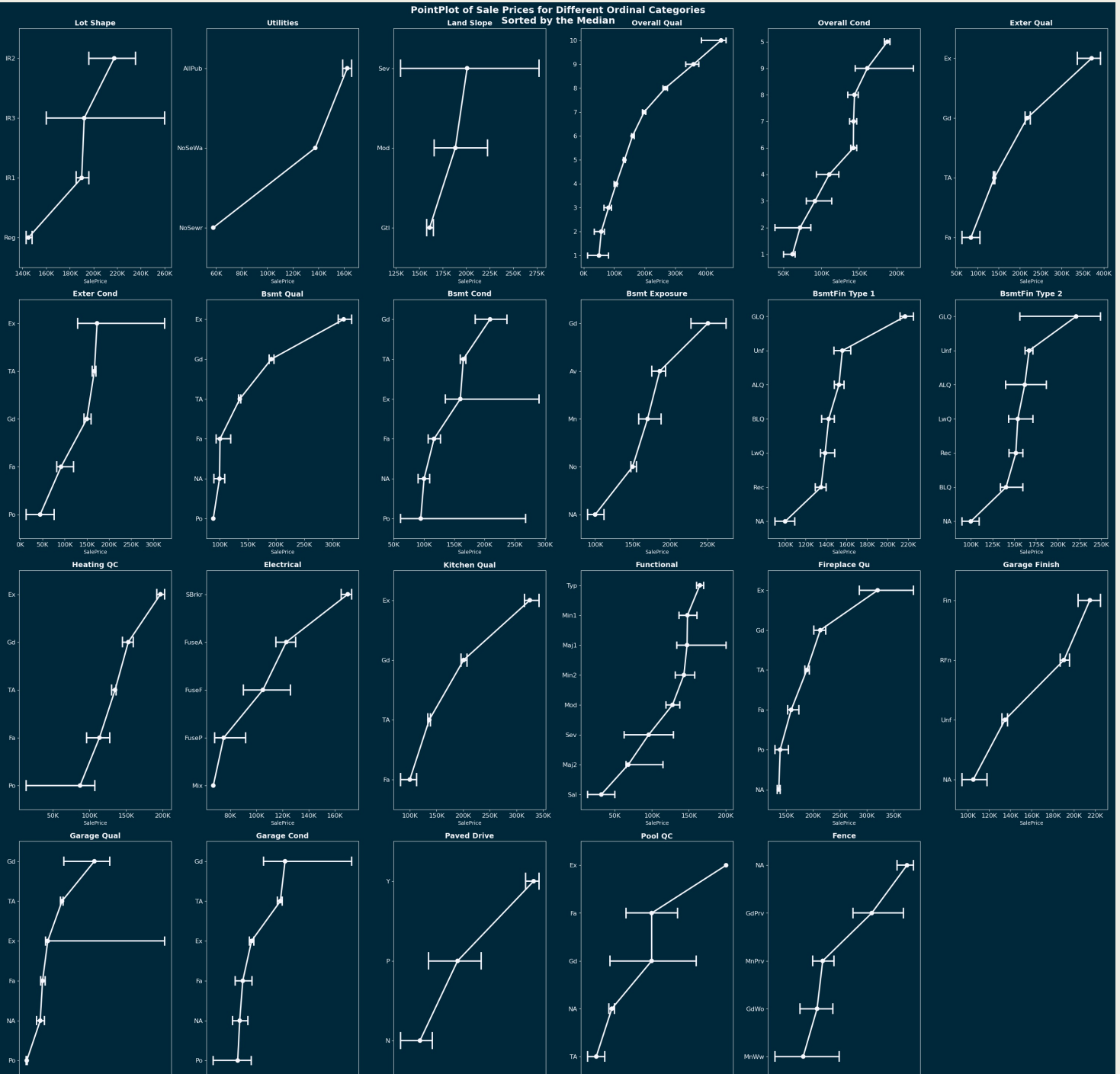
Data Visualizations



Box-Plot of Sale Price for Ordinal Variables

Model Building Workflow

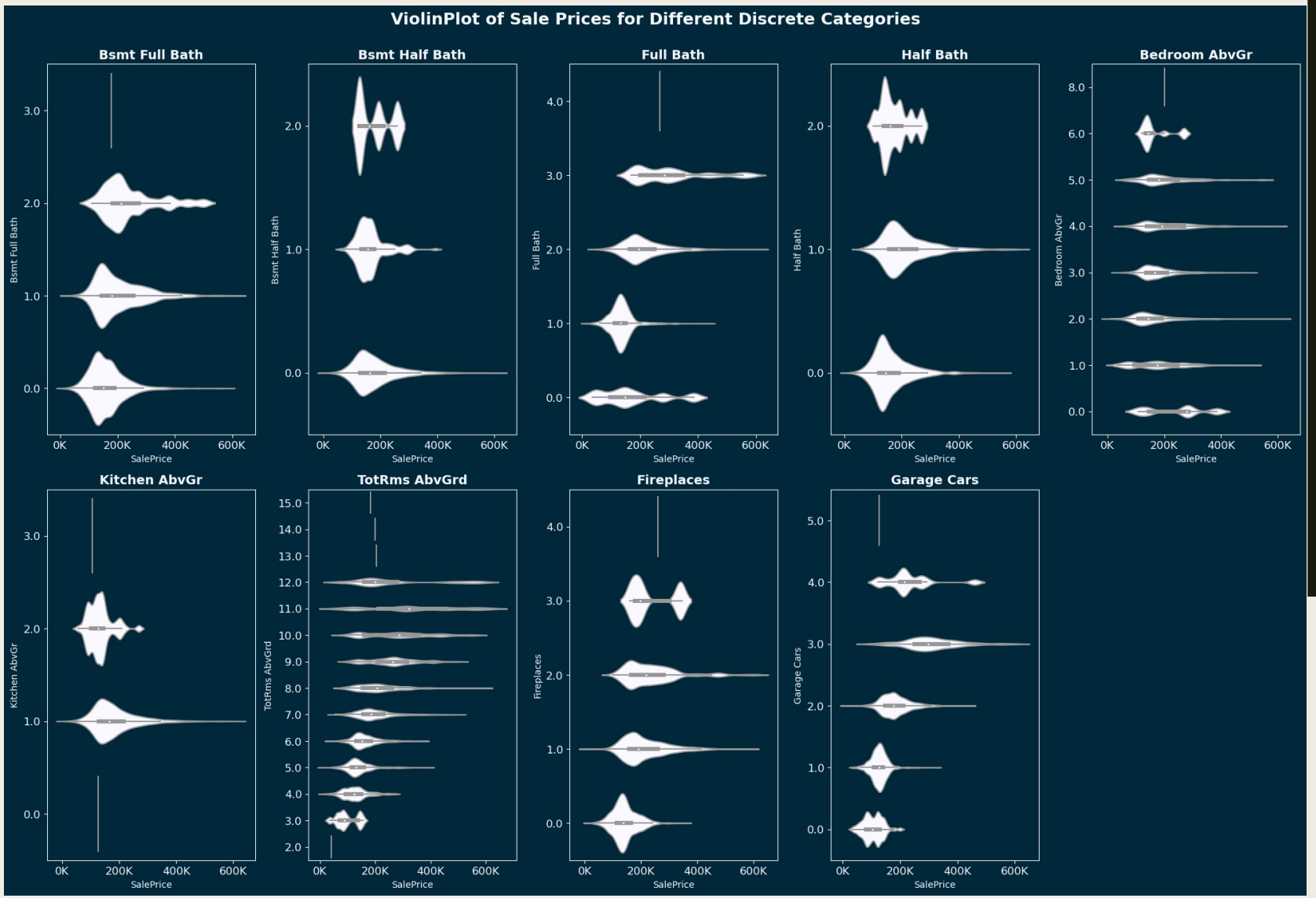
Data Visualizations



Point-Plot of Sale Price for Ordinal Variables

Model Building Workflow

Data Visualizations



Violin-Plot of Sale Price for Discrete Variables

Model Building Workflow

Data Visualizations

1. Feature selection by investigating the relationship between variables using the data visualization generated above.
2. Feature engineering was used to replace and/or reduce the count of existing features.
3. Standardization for Ridge and Lasso
4. Identifying outliers influencing the model performance
5. Iterate through the above steps primarily to minimize the sale price prediction error and get a good variance-bias tradeoff.

Feature Selection and Engineering

Dummies

Bldg Type
Central Air
House Style
Sale Type
Exterior 1st
Neighborhood
MS SubClass
Foundation
Garage Type
Condition 1
MS Zoning
Half Bath
Full Bath
Bsmt Half Bath
Bsmt Full Bath
Bedroom AbvGr
Kitchen AbvGr
Garage Cars
BsmtFin Type 1
Bsmt Exposure
Overall Qual
Kitchen Qual
Heating QC
Paved Drive

Map & Combine Labels

Pool QC
Garage Qual
Fireplace Qu
Bsmt Qual
Garage Finish
Exter Qual

Original

Total Bsmt SF

Gr Liv Area

New

Difference of Year Built and Yr Sold

Difference of Year Remod/Add and Yr Sold

Sum of Wood Deck SF,Open Porch SF,Enclosed Porch,3Ssn Porch,Screen Porch

Binarized Pool Area

36 primary features. 142 total features.

Model Building Workflow

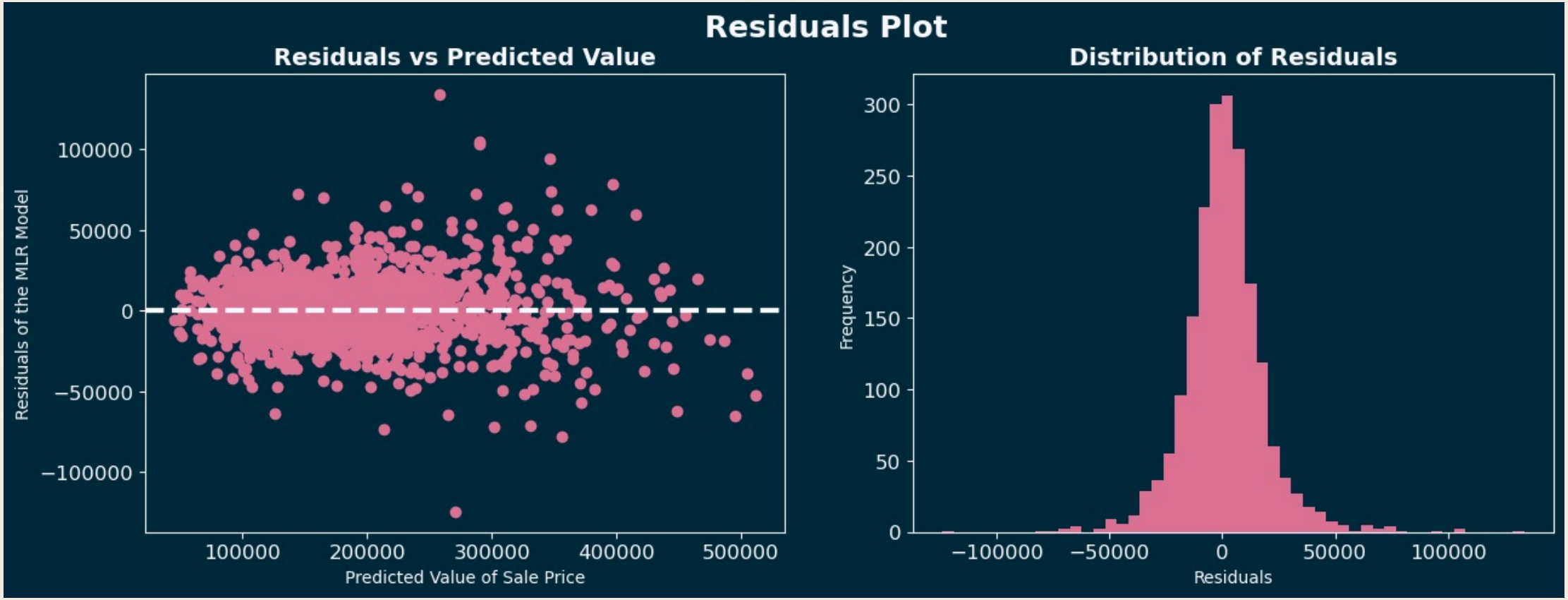
Feature Selection & Engineering

Model Evaluation

- Four linear regression models were developed to predict the sale price
- Sale Price was log-transformed
- Non-regularized and Ridge, Lasso and ElasticNet
- Models were first cross-validated by splitting the train dataset into a train-test split
- After achieving a satisfactory bias-variance trade-off with the best possible RMSE scores, predictions were generated for the test dataset.

Model	Cross-Validated Mean	R2 score (train,test)	RMSE
Linear	1.12	(0.929,0.918)	17,966
Ridge	1.12	(0.929,0.916)	18,043
Lasso	1.12	(0.928,0.916)	18,080
ElasticNet	1.12	(0.929,0.916)	18,081

Since the cross validation was done on log transformed sale price, the RMSE score is finding the square root of the mean of the squared ration between the model values and true values. i.e., if the RMSE were 0.693 (=ln 2) the model values would be roughly a factor of two out on average (in either direction) from the true values in the original (non-log) space.



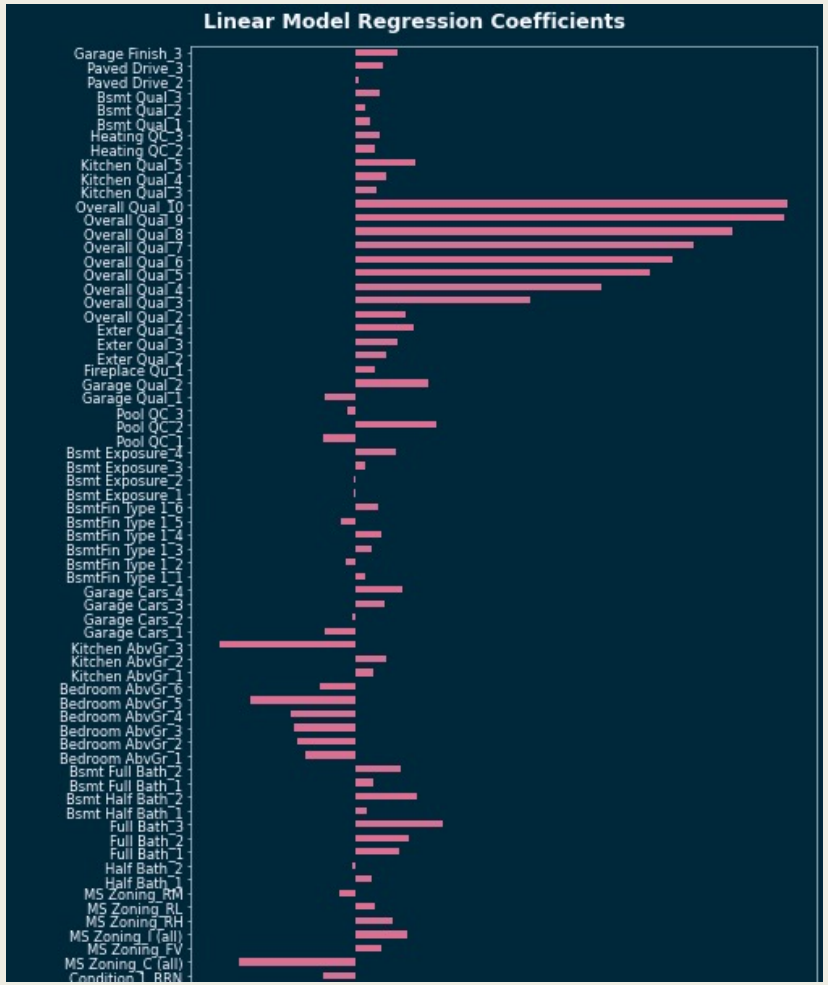
Residuals were investigated for

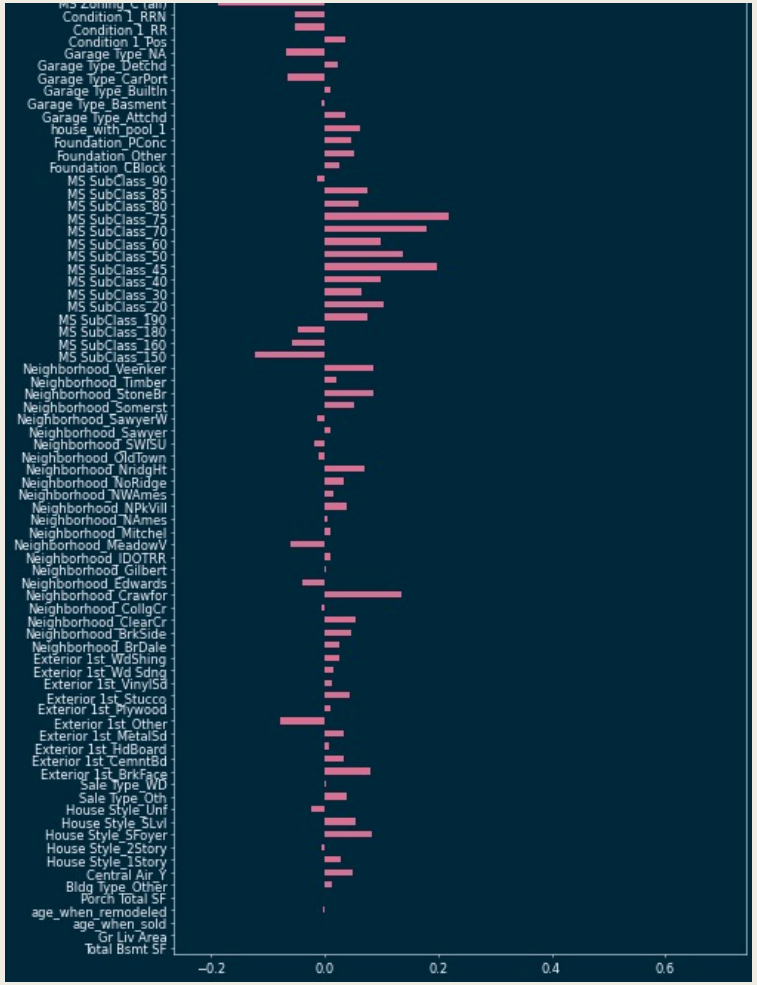
- normality
- homoscedasticity of errors
- independence of observations.

Cross-plot of predicted vs actual sale price to compare against the 45-degree line

Model Building Workflow

Model Evaluation





Linear Model Regression Coefficients

Model Building Workflow

Model Evaluation

ELASTICNET MODEL REGRESSION COEFFICIENTS

Conclusions

- Ames housing dataset is used to train and fit a linear regression model to estimate sale price of a house
- The RMSE score for this model is 18,000
- The model is validated for the linear model assumptions
- Overall quality is the biggest positive contributor to the sale price and not having a garage is the biggest negative contributor