Project 2:
Ames Housing Data

Presented by:
Kara | Lee Mei | Wei Hua | Wenna

Background

Investment firm that focuses on investing in residential properties

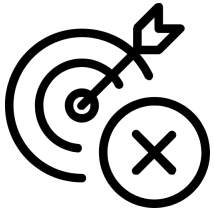
Newly hired
Data Science team

Tasked to use the Ames housing dataset to build a model that predicts the price of houses in the city

Problem Statement

Housing prices are currently estimated manually by experts.
A team gathers up-to-date information about a residential property, and estimates the sale price by manually assessing key characteristics that will most likely influence it.

The process is costly and time-consuming, and their estimates are not great



The model's output (a prediction of a house sale price) wil be fed to another machine learning system, along with many other signals.

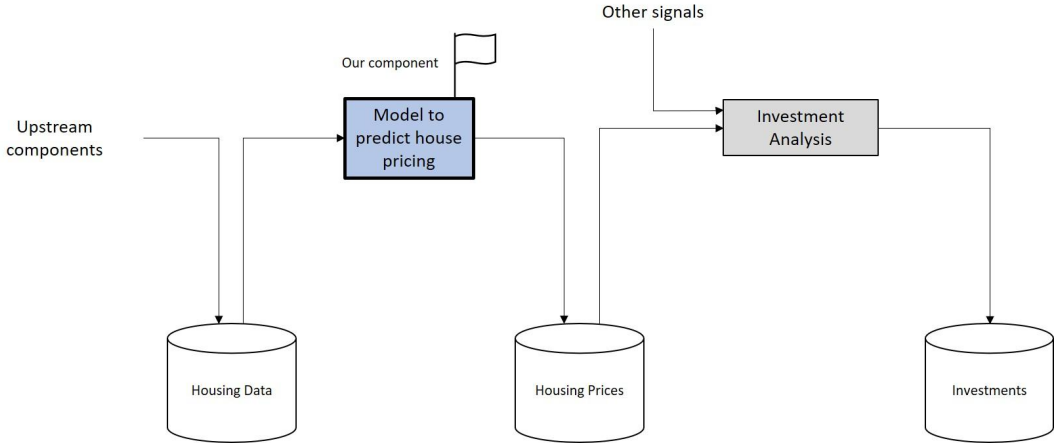
Stakeholders

Primary

Team of experts who are manually estimating housing prices

Secondary

Management team of the housing investing firm

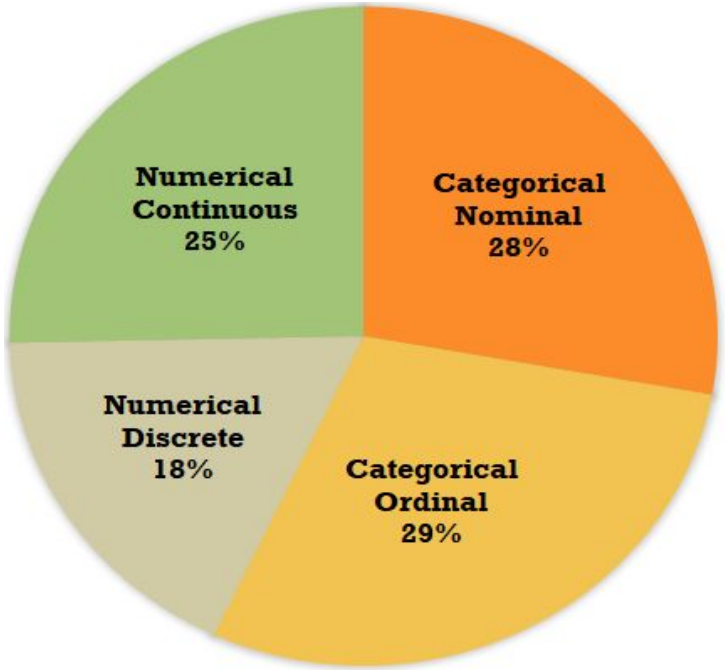


A machine learning pipeline for residential property investments

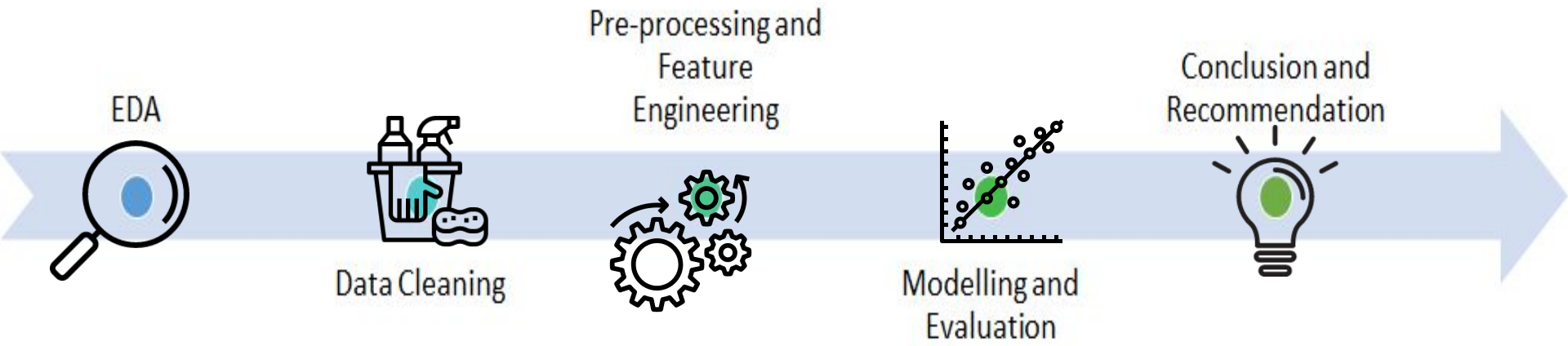
Provided Dataset

The Ames housing dataset contains information on the value and sale price of individual residential properties sold in Ames, Iowa from 2006 to 2010.

The dataset contains information on 2929 observations, and 81 variables.



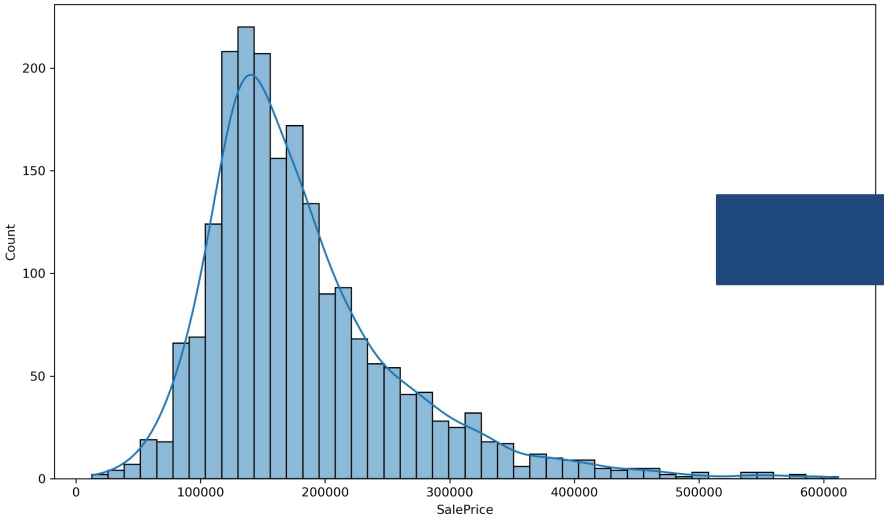
Workflow


EDA
Data Cleaning
Pre-processing and Feature Engineering
Modelling and Evaluation
Conclusion and Recommendation

EDA

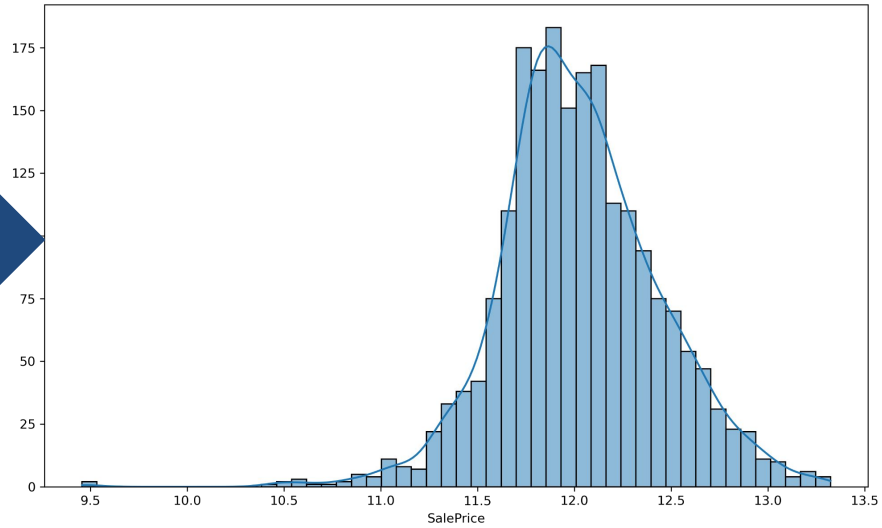
Target Variable: Sale Price

Before:



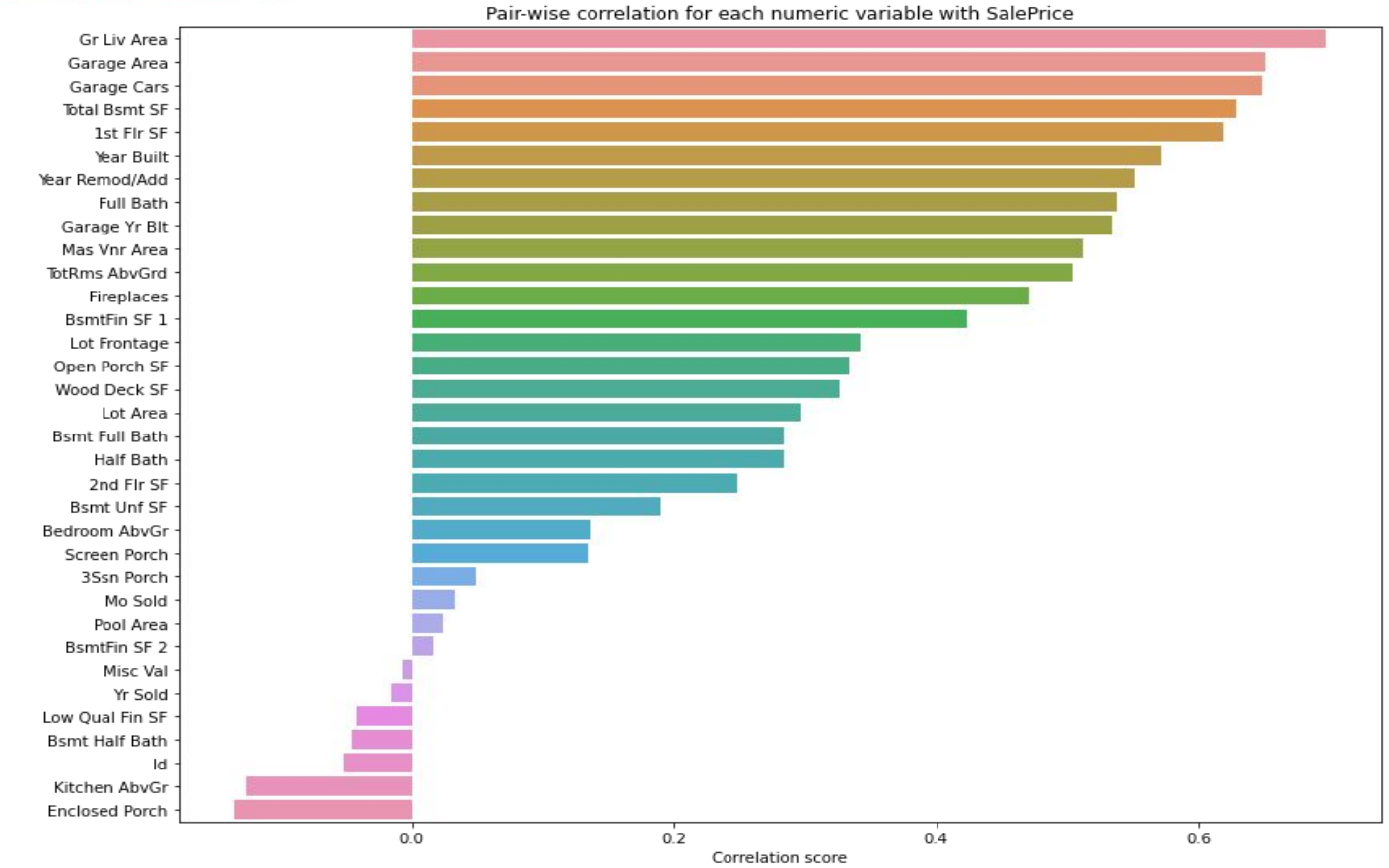
The SalePrice target variable is **positively skewed** (skewed right): its right tail is longer and most of the distribution is at the left.

After:



We undertook a **log transformation** of the highly skewed SalePrice target variable. The histogram of log sale prices **appears more symmetrical**, with less extreme values.

The most important numerical features

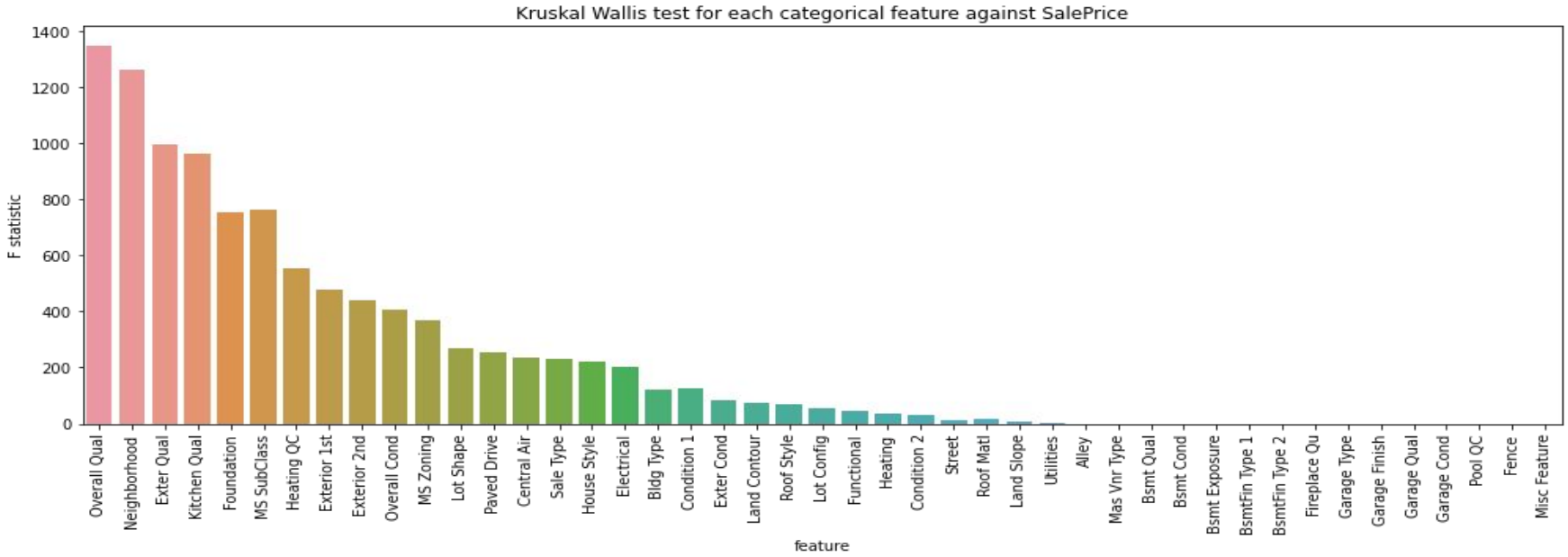


Overall, we can see that variables associated to:

- Home Size
- Usable space &
- Age

of a house are critical in influencing sale price.

The most important categorical features



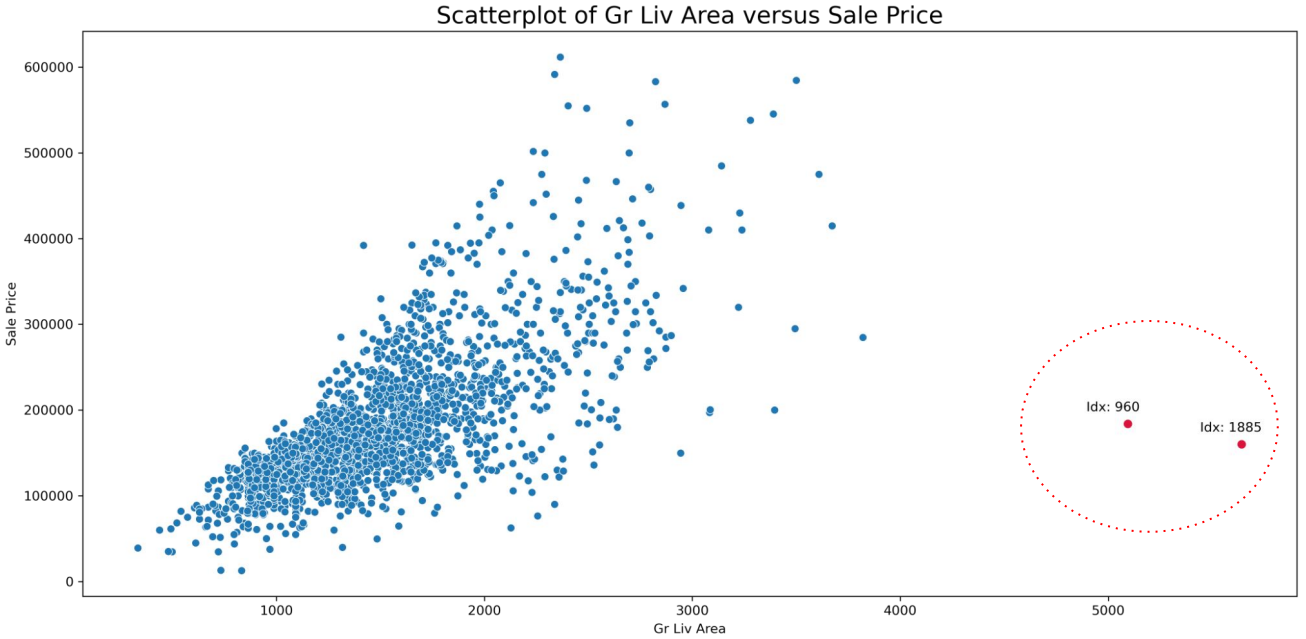
5 of those most important categorical features are: Overall Qual, Neighborhood, Exter Qual, Bsmt Qual, and Kitchen Qual. We noticed that the **quality variables are important** in influencing sale price. We can also confirm the real estate mantra of **"location, location, location"** as Neighborhood appears to have the second greatest influence on sale price.

Data Cleaning

Outliers

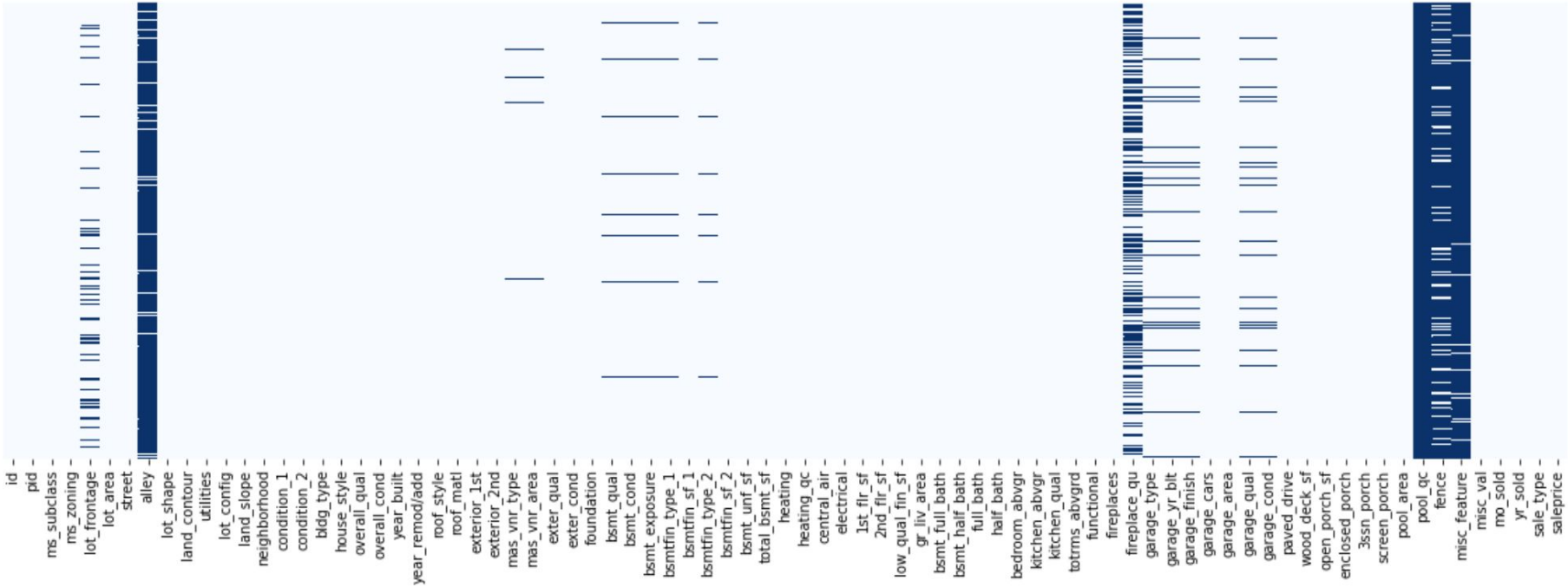
Finding:

There are a number of outliers within the data.



For example, these are properties that have more than 4000 square feet.
The two houses seemed to be unusual sales (very large houses priced relatively appropriately).

Missing Data



Most of the missing data corresponds to the absence of a feature. For example, the PoolQC variable has the most missing values. The high number of missing values here makes sense as in reality only a small proportion of houses will have a pool. These values are imputed as 0 or "None" depending on its data type. For houses with a feature present but have missing values, the strategy is to treat these values using imputation techniques.

Preprocessing and Feature Engineering

Feature Engineering

Created new input features are created from existing features

Examples:

Total Bathrooms:

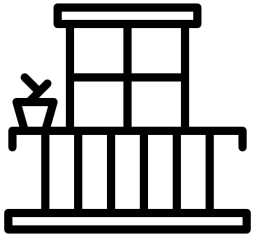
Total number of bathrooms in a house

Full Bath + 0.5*Half Bath +
Bsmt Full Bath + 0.5 * Bsmt Half Bath

Total Outdoor Area:

Total square feet of all outdoor areas

Wood Deck SF + Open Porch SF + Enclosed Porch + 3 Season Porch + Screen Porch



Pre-processing

Some of the pre-processing steps undertaken are:

- Dropping of highly correlated variables to reduce multicollinearity
- Normalising numeric predictors
- Label encoding of ordinal features
- One-hot encoding of nominal features
- Feature scaling

Modelling and Evaluation

Model Evaluation

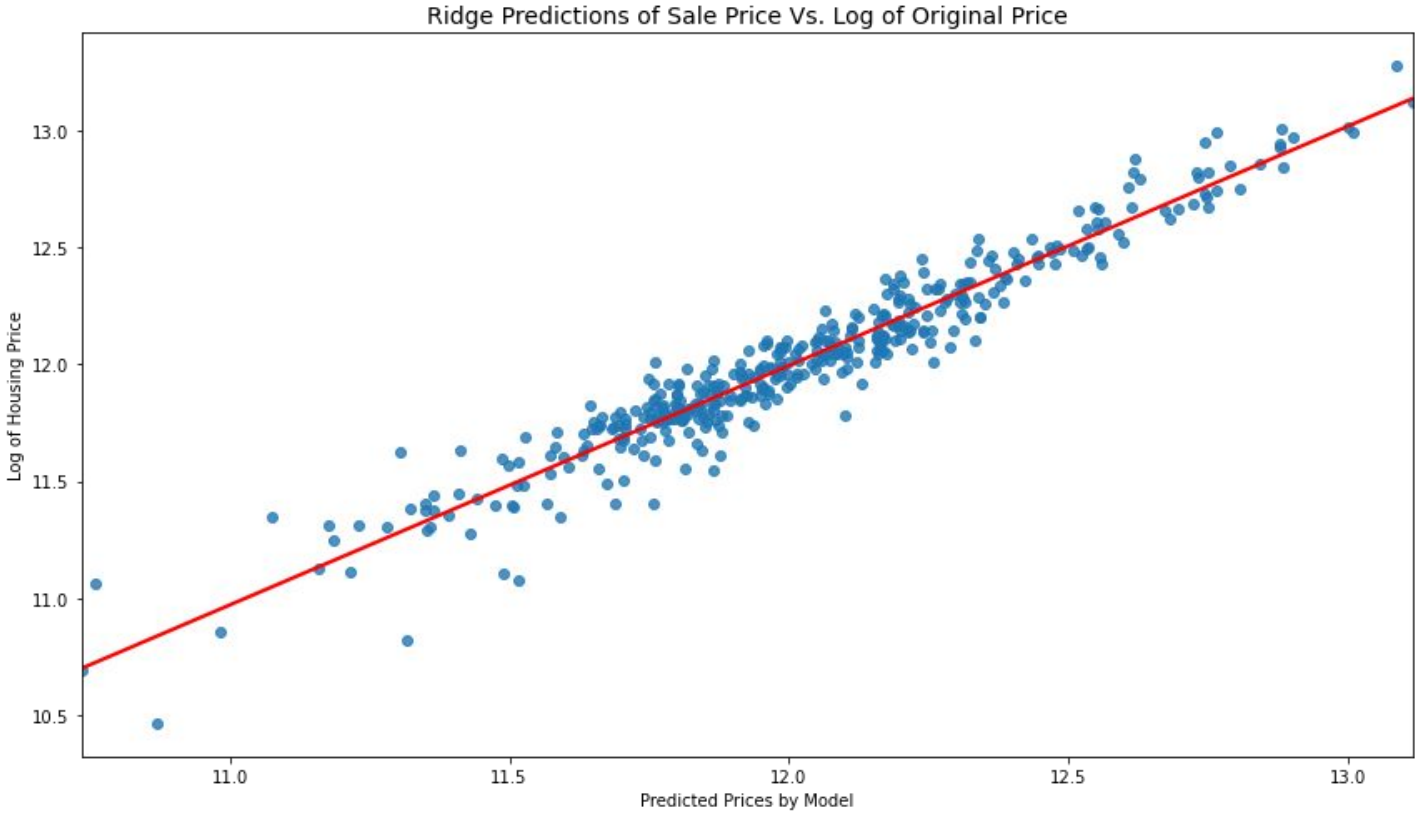
Goal: To have the RMSE score as close to 0 as possible

Evaluation metric of 3 regression algorithms:

Model	Mean RMSE score of 5 runs (log scale)
Linear Regression	18268020762.45
Ridge Regression	0.1232
Lasso Regression	0.1241

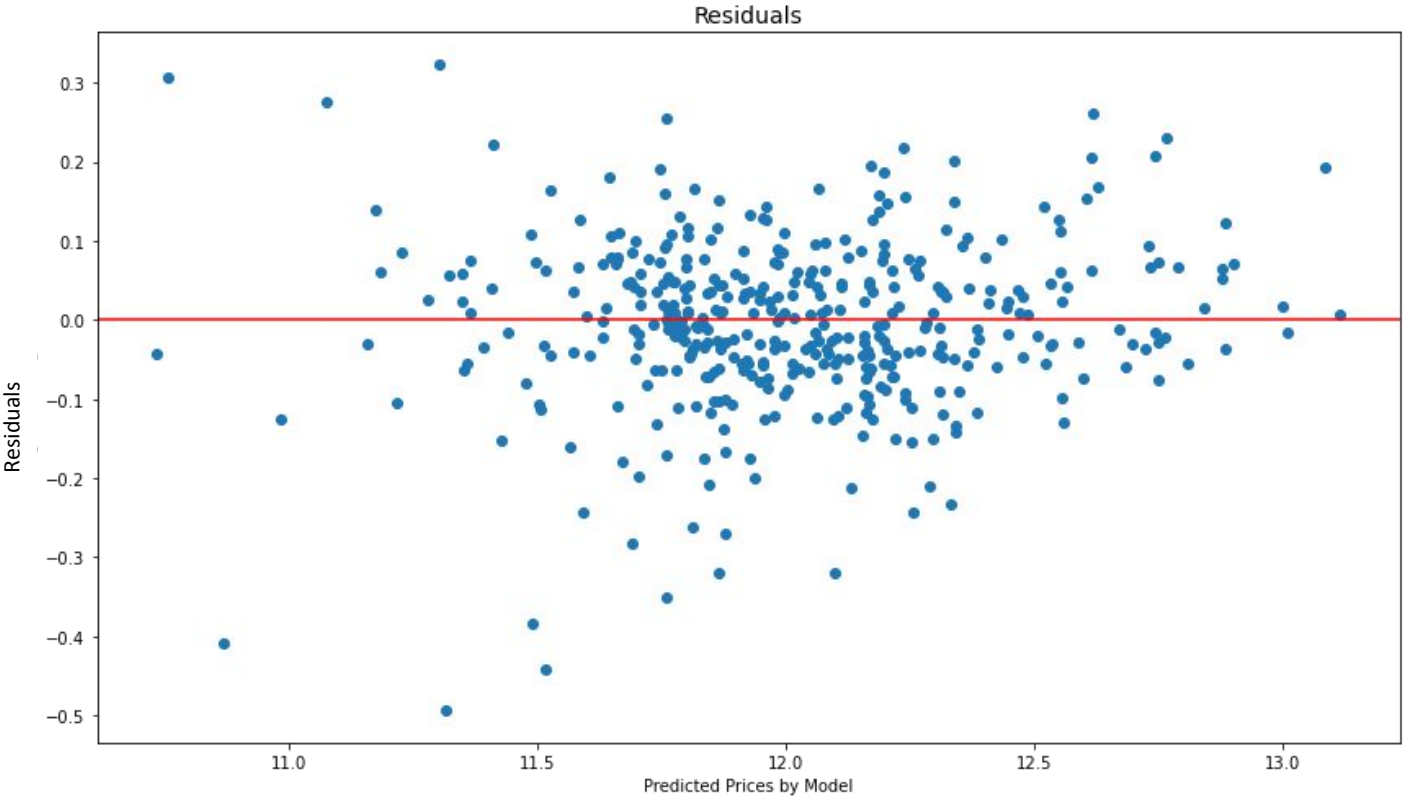
After applying 5-folds cross validation, it is clear that Ridge Regression model has the best performance in terms of RMSE score.

Production Model



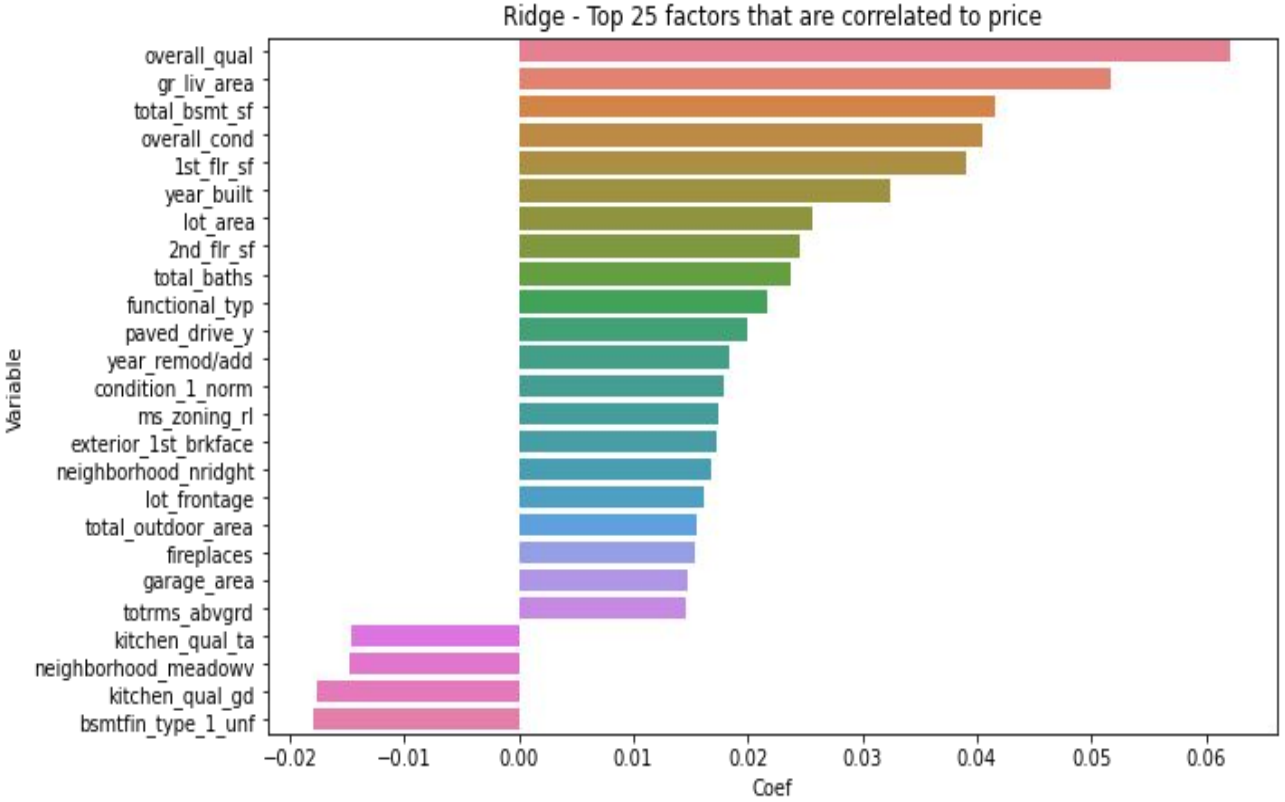
We can see that the line of best fit passes through most of the points

Production Model



The residual plot shows a fairly random pattern - supports the assumption of linear model. We are able to see a consistent variance between our low predictions and our high predictions (homoscedasticity).

Important features for the model



From the graph, we can tell that the features with the Top 10 correlation are related to
2 main groups - Size and Condition of the house.

1) Size
- Ground Living Area, Total Basement SF, 1st Floor SF, Lot Area, 2nd Floor SF, Total Baths

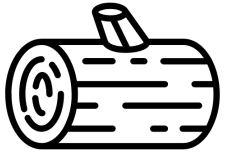
2) Condition
- Overall Quality, Overall Condition, Year Built, Home Functionality (Typical)

Conclusion & Recommendations

Conclusion

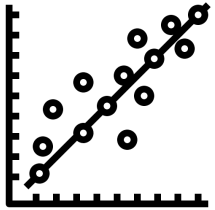
Based on the insights from our exploratory data analysis (EDA), we zoomed in on the various features that are likely to influence SalePrice.

There followed by data cleaning, preprocessing and feature engineering to prepare our data for model training.



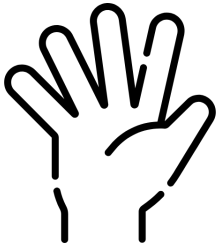
We also undertook a log transformation of the highly skewed SalePrice target variable.

Conclusion

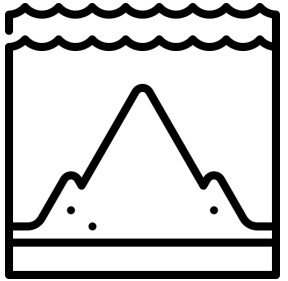


We tried to predict the SalePrice of a house with regression models. The engineered data was then run through three regression algorithms:

- Linear
- Lasso
- Ridge



After applying 5-folds cross validation, it is clear that Ridge Regression model has the best performance in terms of RMSE score.



Hence, we chose the Ridge Regression to create our production model.

Recommendations

Based on our research findings, we would recommend the following steps going forward:

Primary

Team of experts who are manually estimating housing prices

1. Model eliminates the need for manual calculation currently performed, and will help to free up resources for them to focus their expertise on more productive tasks
2. Use the insights from the model to identify potential features (e.g. assorted property size variables, namely Ground Living Area, Total Basement SF) for improvement to increase Sale Price

Secondary

Management team of the housing investing firm

1. Model helps identifies key features that increase Sale Price to boost the firm's portfolio returns
2. Offers better judgement/insight for potential real estate investments

Limitations

Location specific:

- Although our production model generalises well to the houses sold in Ames, Iowa, it probably will not generalise well to other cities, given that each city tends to differ greatly in terms of geographical features, climatic conditions, and cultural preferences.

Outdated data:

- Another point to note is that this model works well for houses sold from 2006 to 2010. Housing prices would have varied over the years, especially after the financial crisis in 2008 and the recent covid-19 crisis. Hence, our model would need to be retrained using more recent data.

Nature of transactions:

- The model assumes sale and purchase are made by willing buyers and willing sellers. Hence, the model may not be suitable for auction sale or forced sale or quick sale.

Thank you