**BUDT 758B Final Project Summary:**

**Ames Housing Price Prediction**

**Group Name: Data Bang Theory**

**Group Members: Lei Han, Xiaoyou Zhou, Zhonghao (Allen) Li, Yutian Luo, Siyuan Luo, Jinxin Hou**

**Background:**

Ask a home buyer to describe their dream house, and they will probably come up with many criterias, based on which, we want to help the home buyers to predict their dream house price.

We obtained our dataset from Kaggle. com, which includes 1 response variable (sales price) and 79 explanatory variables describing residential homes in Ames, Iowa.

**Objective Goal:**

* Our goal is to predict house prices with labels and attributes of Ames Housing dataset.
* Our models are evaluated on the Root-Mean-Squared-Error (RMSE) between the log of the SalePrice predicted by our model, and the log of the actual SalePrice.

**Model Key Features:**

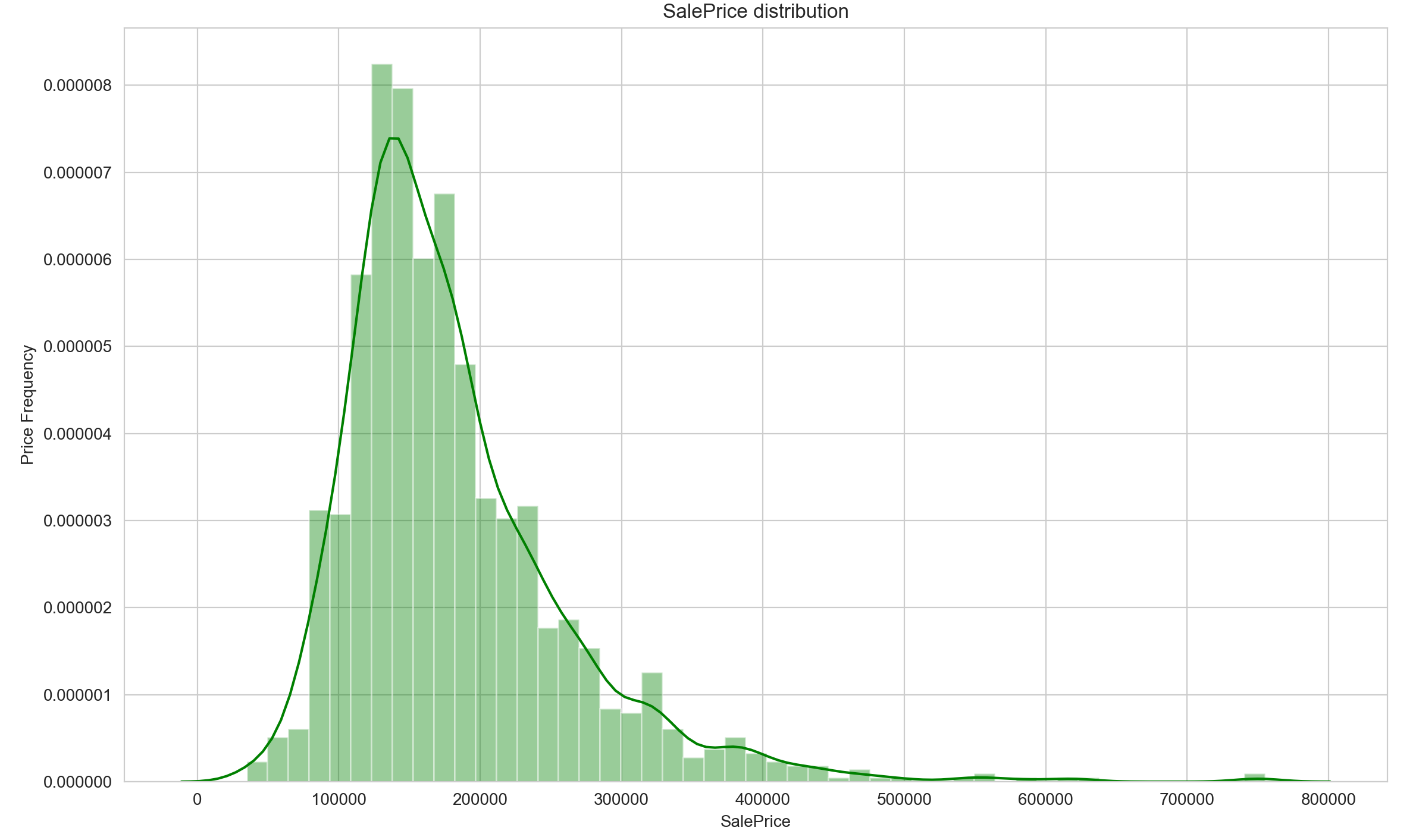
* **Cross Validation:** Using 12-fold cross-validation
* **Models:** On each run of cross-validation we fit 3 models (ridge, random forest and xgboost)

**Methods:**

The standard process is followed to predict house price: EDA, Data Preparation, Model Selection and Prediction.

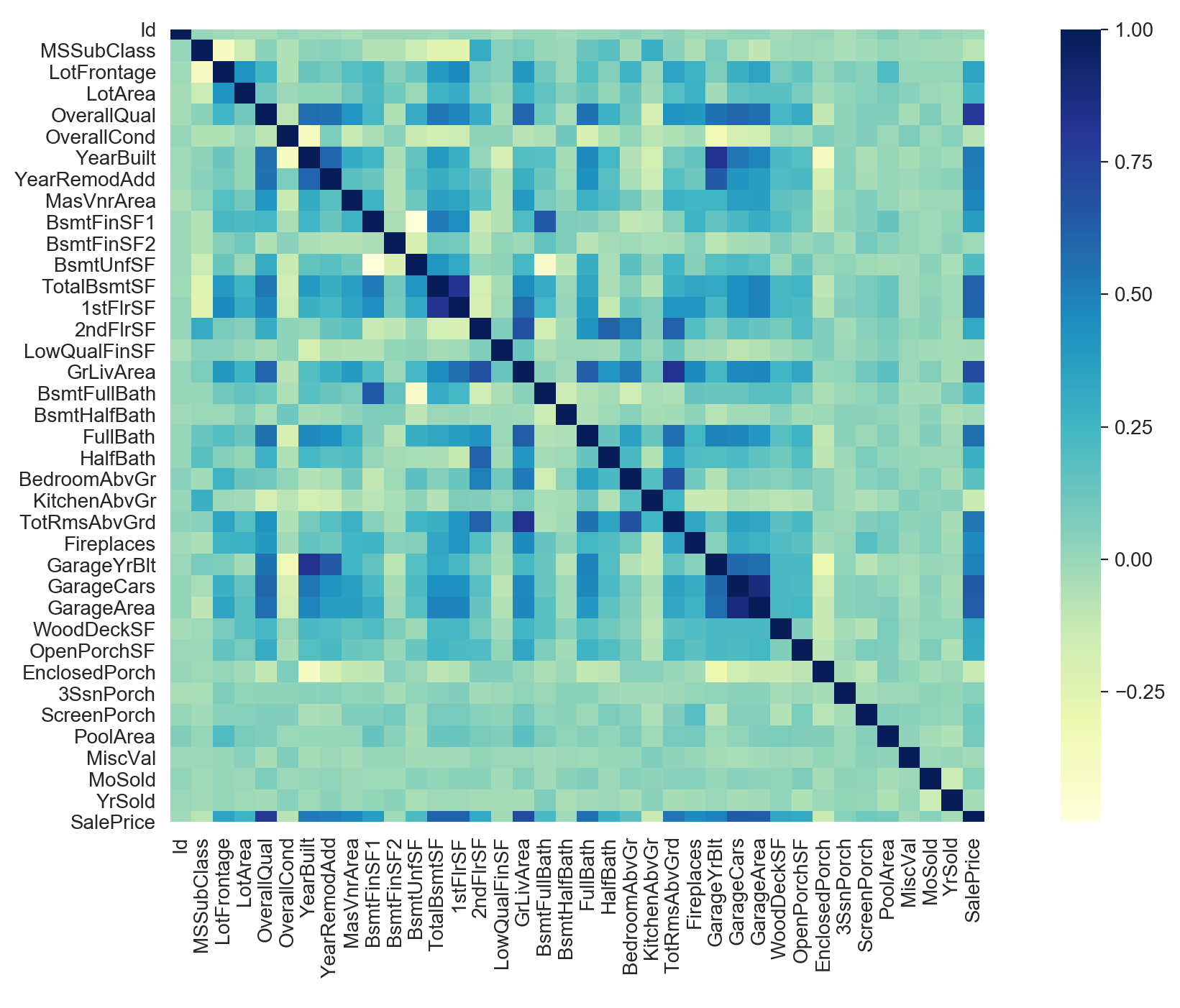
1. **EDA**

Fig 1.1:Distribution of “Sale Price”



With the histogram, we can view the skewness of the label distribution, which is useful for us to do the following data preparation. And from the plot we can see that sale price obeys right skewed distribution, which aligns with our common sense.

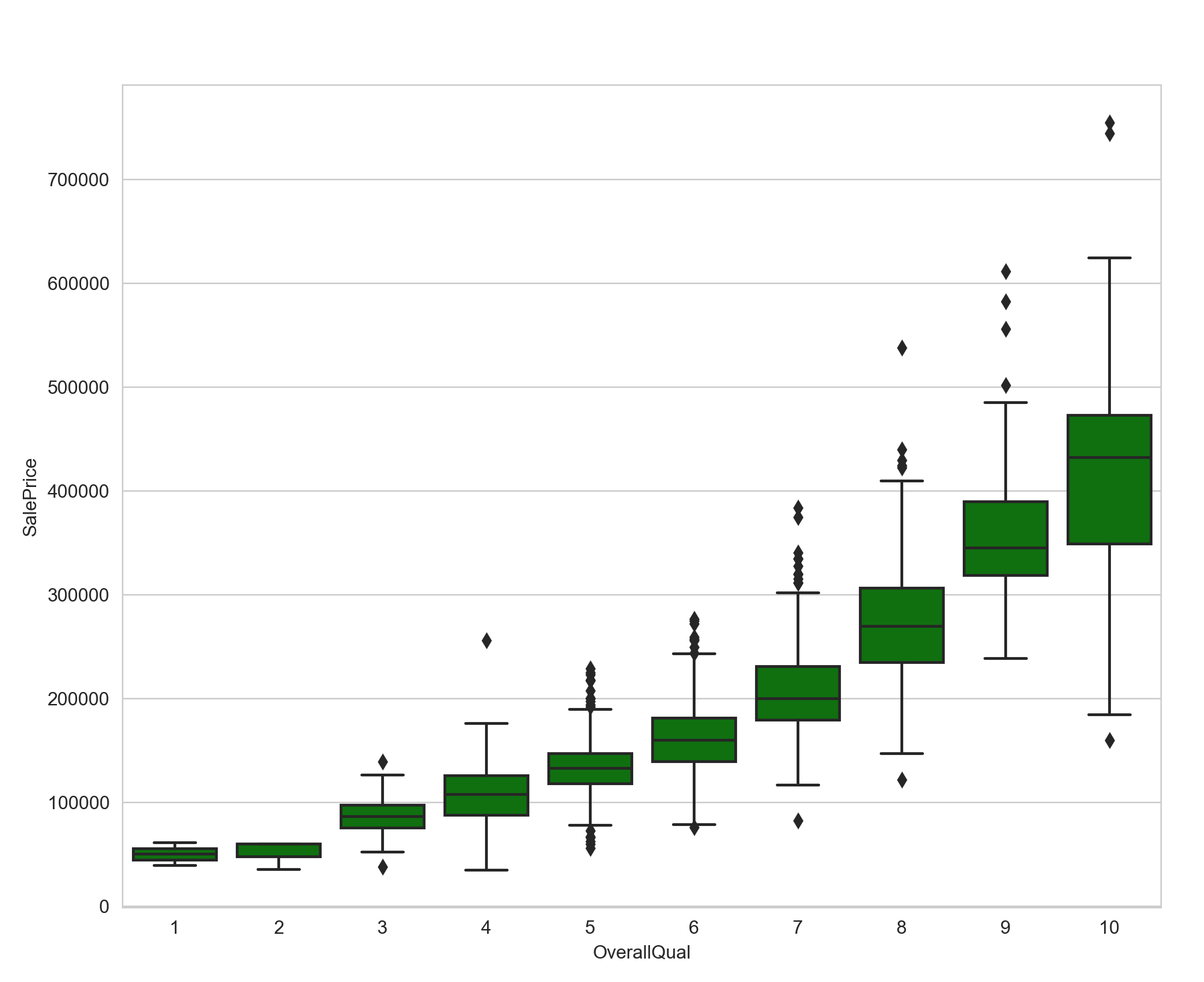
Fig 1.2:Relationship Between Attributes and “Sale Price”



We can easily spot all the relationships between attributes and house price from the last row of the heatmap, where we can find the attributes are highly correlated to house price. Then we can focus on these attributes to do the feature engineering and modeling.

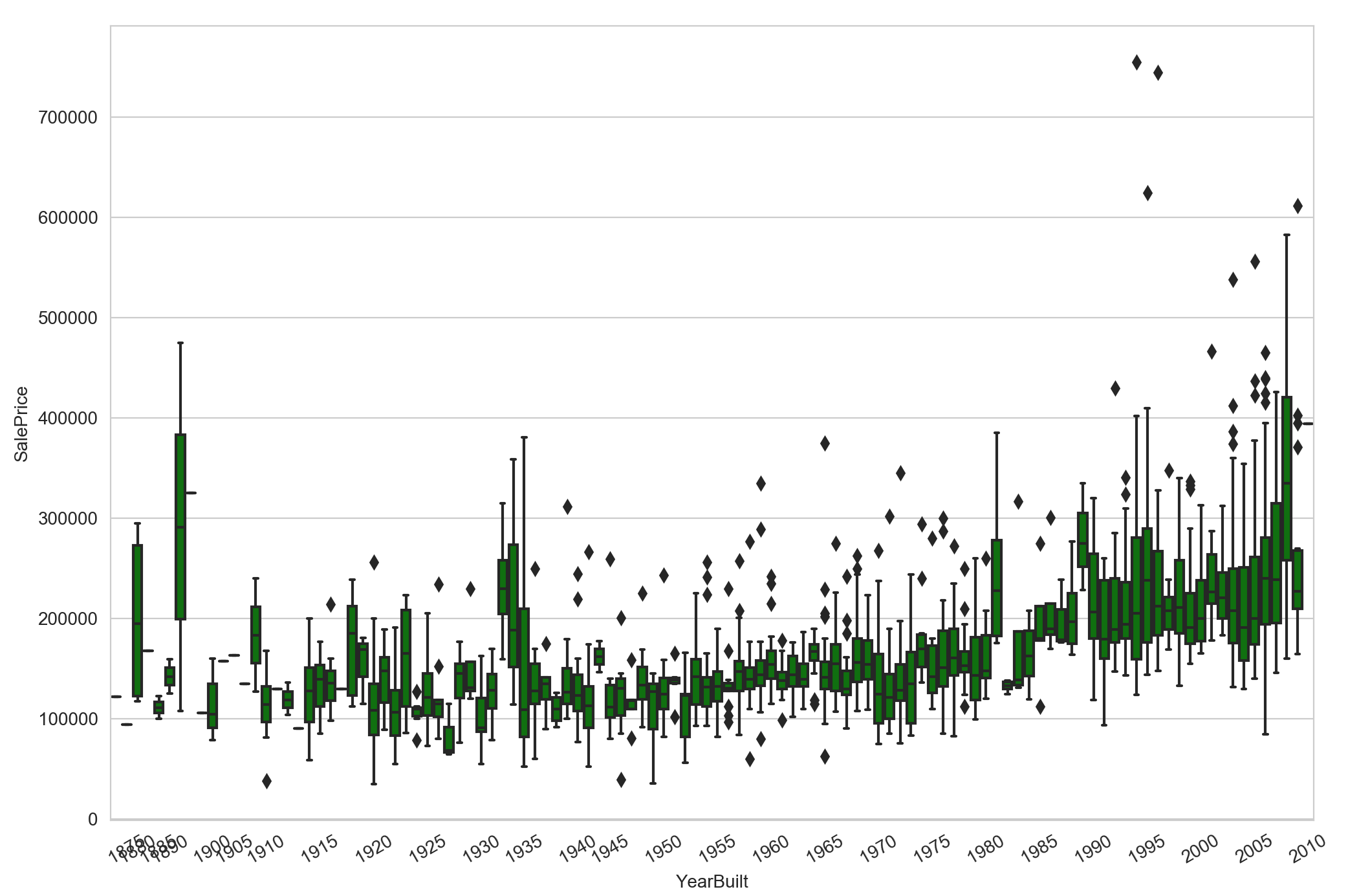
And from the heatmap we can see that "GarageYrBlt" and "YearBuilt" are highly correlated, "GarageCars" and "GarageArea" , "TotalBsmtSF" and "1stFlrSF" are also highly correlated. Therefore, when building our model, we will delete one feature of each pair.

Fig 1.3*:* Relationship Between “Overall Quality” and “Sale Price”



This graph suggests that Overall Quality is positively related to the house price. Therefore, Overall Quality is an important feature and we can use it for predictive modeling.

Fig 1.4*:* Relationship Between “Year Built” and “Sale Price”



From the graph we can see the price trend of the houses built in different years.

1. **Data Preparation:**

* Data Transformation

- Convert the data type of some columns to appropriate ones

* Data Cleaning
* Drop unnecessary features which have more than 90% null values
* Fill the NAs in numerical features with mode, mean based on the data or 0
* Fill the NAs in categorical features with reasonable values according to the data description or 0
* Data Normalization

- Normalize all numeric features

- Transform “sales price” by log(1+x)

- Get rid of any duplicate data

1. **Feature Engineering:**

* Create some new features like 'Total\_Home\_Quality' which can represent the feature of the house
* Create dummy variables to represent whether the house has a pool, garage, basement or not

1. **Models:**

We chose and compared outcomes among 3 models: XGBoost Model, Ridge Model, and Random Forest Model for the following reasons. And we used 12-fold cross-validation to validate the model result.

* XGBoost

- As one of most popular predictive models stemming from Gradient Boost Decision Tree, refined its performance by using Taylor's Formula on generic loss function, has proven its performance and generalization ability, and is applied by many data scientists on the top of leaderboards in Kaggle competition.

* Ridge Model

- With adding a L2 penalty on the loss function of linear and logistic regression, is simple as well as computational efficient dealing with enormous data.

* Random Forest Model

- Use bagging as well as sampling dimensions to build up a bunch of independent decision trees, and can be deployed using multithreading computation, which proves its shrinking efficiency in the loss function. It performs exceptional dealing with high-dimensional data, and it comes with the ability to judge importance (entropy) among dimensions.

1. **Outcome:**

We got 4 main plots and several findings:

* “Sale Price” distribution histogram

- Sale price obeys right skewed distribution

* Variables correlation heatmap

-"GarageYrBlt" and "YearBuilt" are highly correlated; "GarageCars" and "GarageArea" , "TotalBsmtSF" and "1stFlrSF" are also highly correlated

- Preparation for modeling

* “Overall Quality” to “Sale Price” boxplot

- When the overall quality increases, the sale price will increase;

- When the overall quality increases, the range of sale price will increase as well;

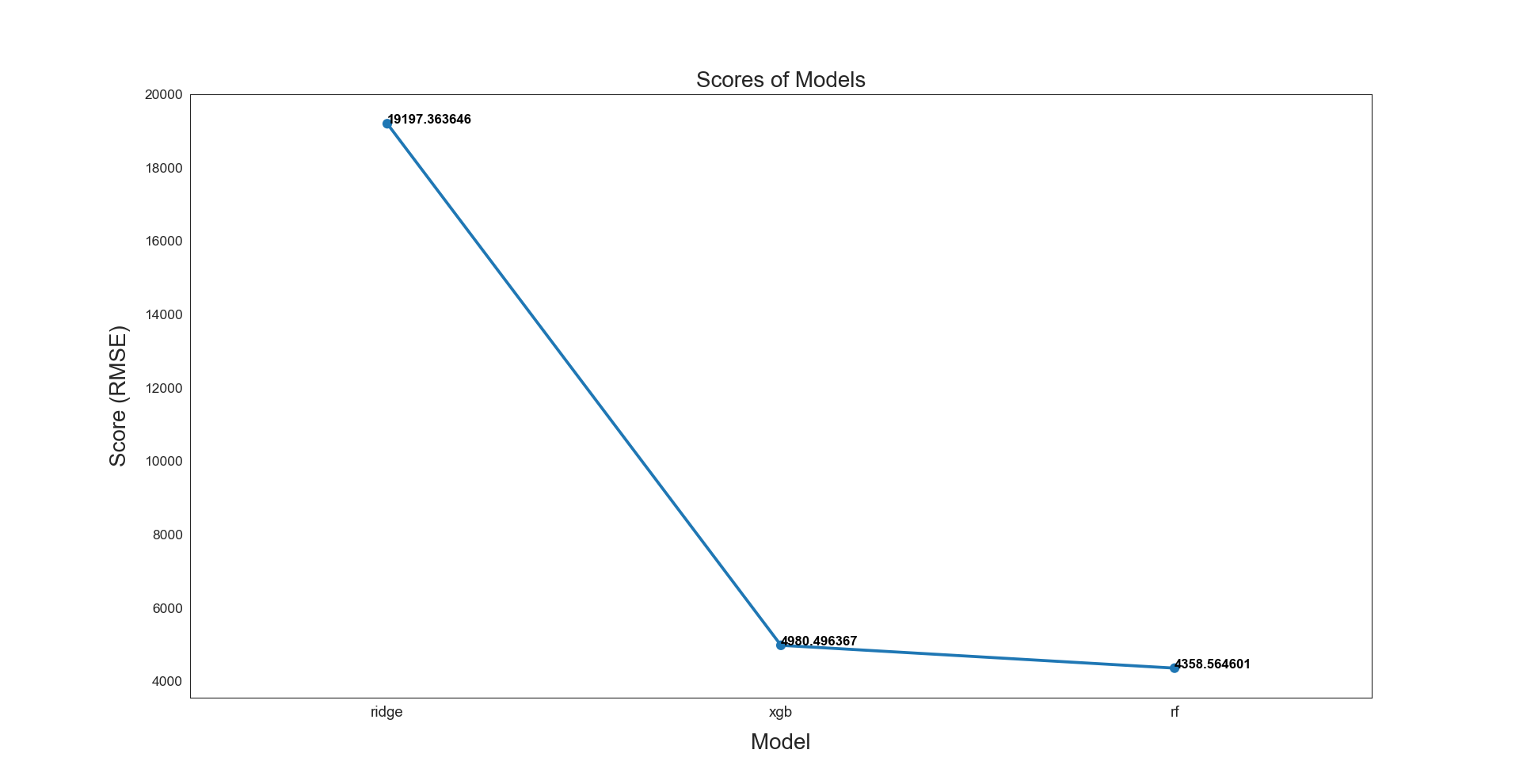
- Almost all outliers are upward anomalies (above the upper bound), indicating that all alternatives that are not following the mainstream are high prices, not the floor prices. The sell side is the dominant side in the real estate market.

* “Year Built” to “Sale Price” boxplot

- The sale price is increasing steadily over the years. However, some old houses might have higher value.

- Explained reasons for why there are 3 sudden increase points and 3 sudden decrease points.

1. **Conclusion:**



After EDA, feature engineering, and modeling, we used the loss function RMSE to gauge all the models. According to the model performance, we decided to use the Random Forest model to predict the housing prices since it has the lowest RMSE score among the three models we used.