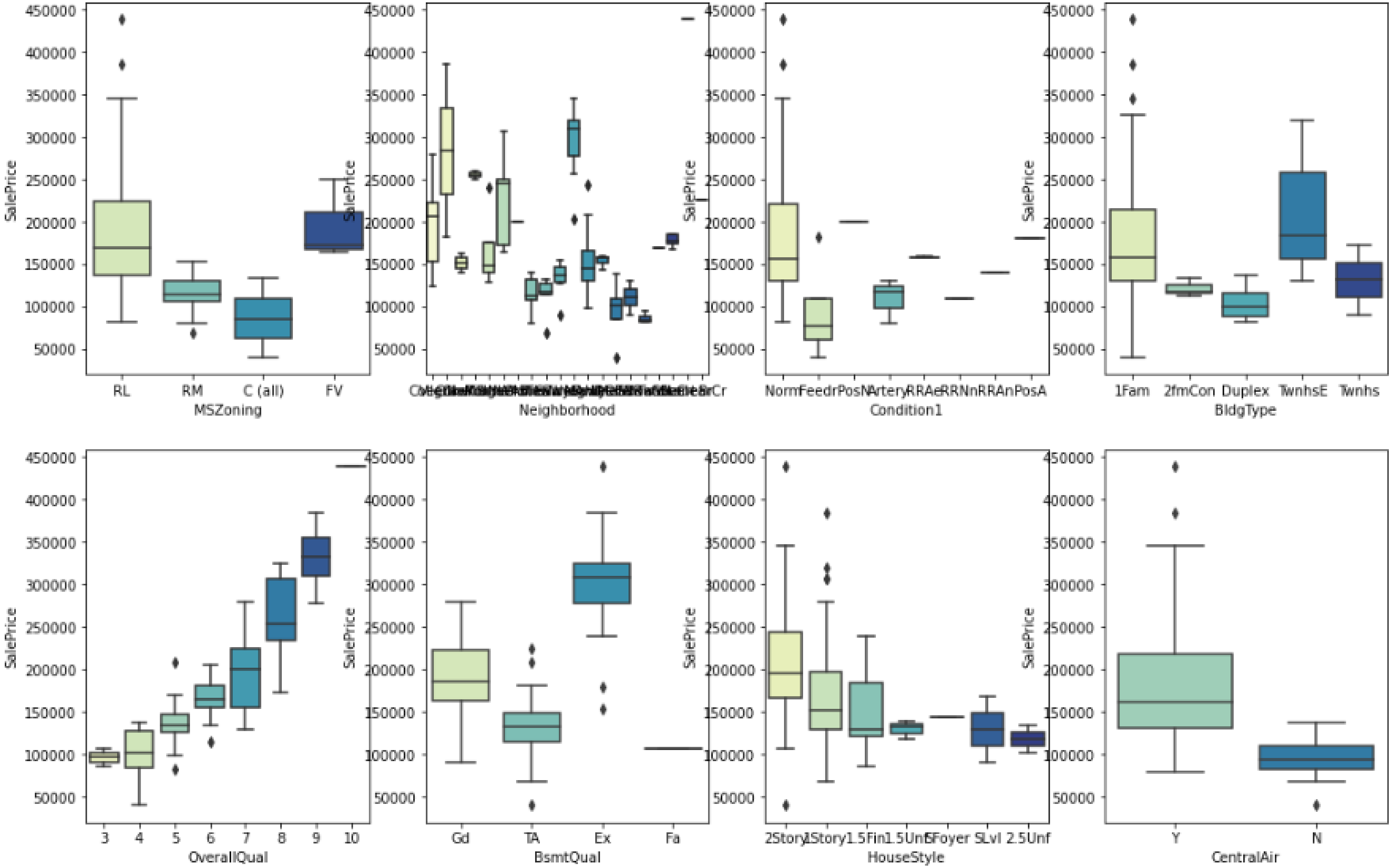
Introduction

- Import historical sale data
- Clean data
- Analyze data
- Train prediction model
- Test model
- Conclude and next steps

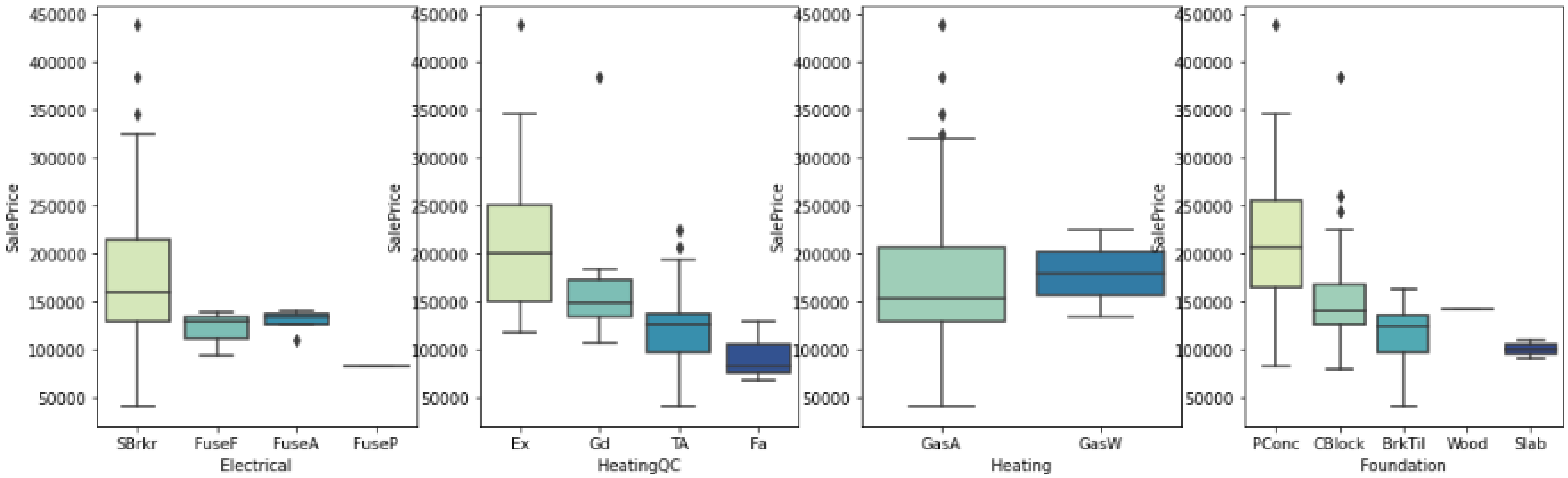
The Data

- 89 columns
- 1000 lines
- Numeric and text values

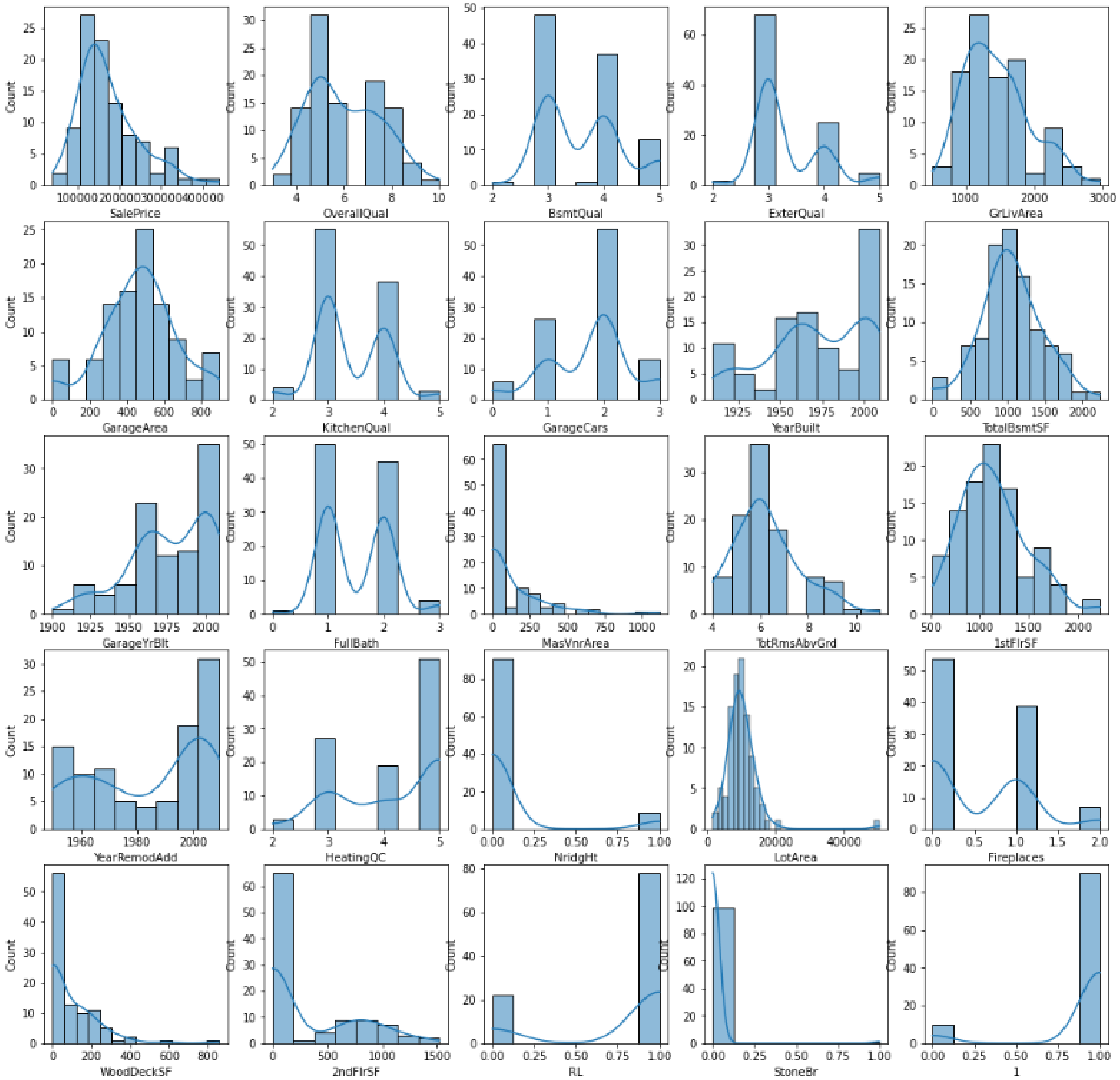
Data Exploration



Data Exploration



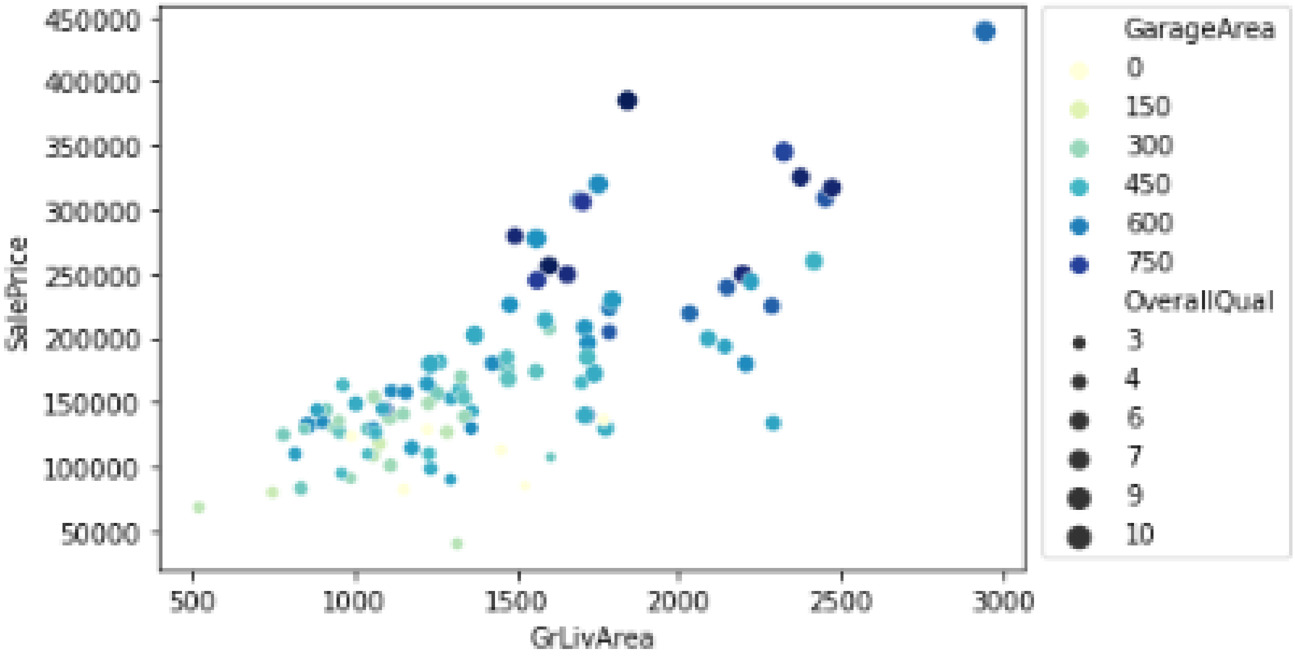
Data Exploration



Not all the variables are normally distributed

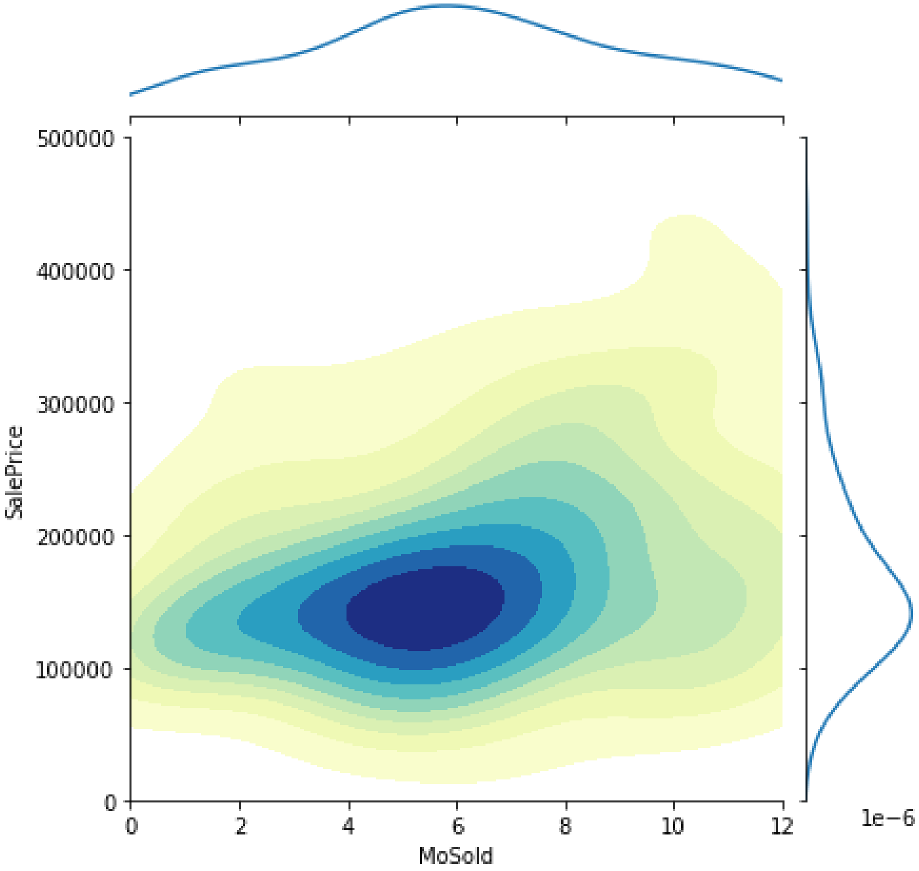
Data Exploration

This plot shows that the sale price is proportional to the ground living area. When the price is overestimated, it is because either the overall quality is high or there is additional space coming from the garage.



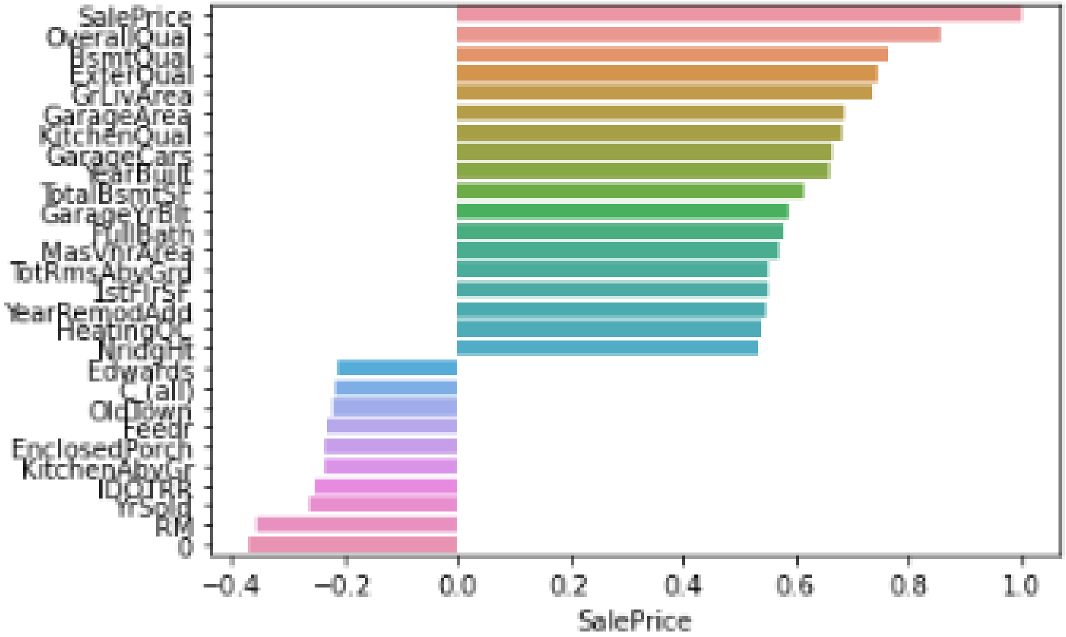
Data Preparation

It can be seen that from January to March, the market is very quiet, not a lot of sales and little SalePrice. The highest number of sales are around June. The biggest SalePrice will happen around November



Data Preparation

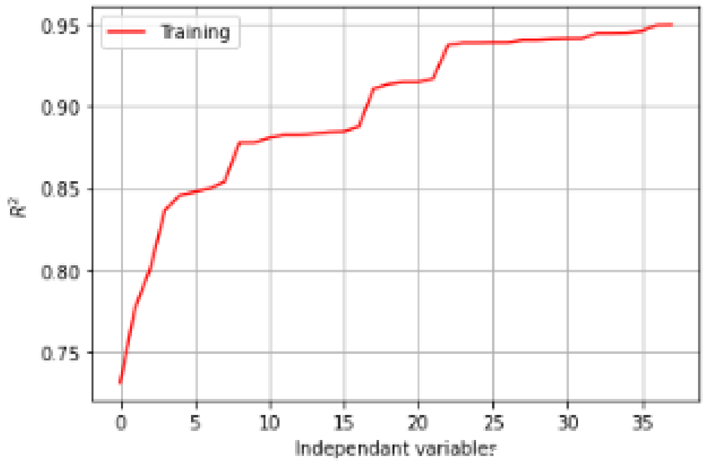
Keep only numeric variable and filter only the variables that are the most correlated (positively and negatively) to the SalePrice variable

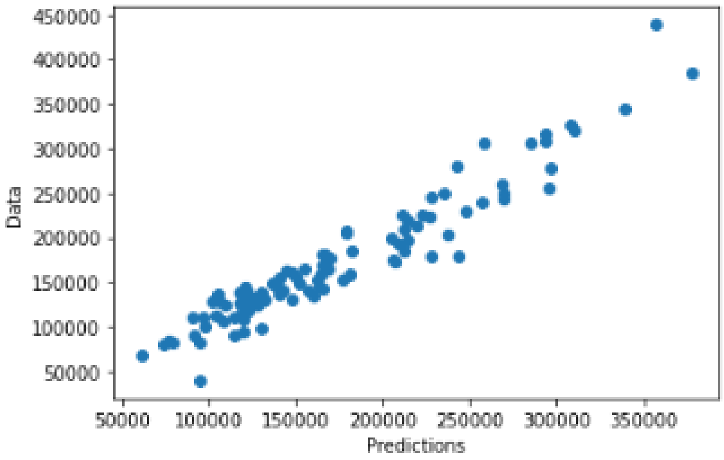


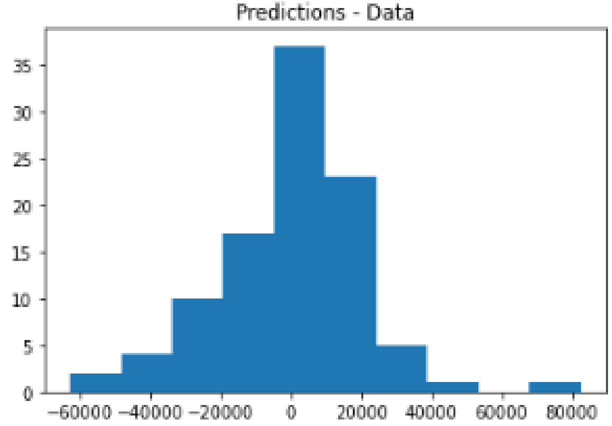
*Please add slides as required.

Correlation

A for loop ranging from 1 to 40 has been created to iterate on the number of variable included to train the linear regression model. The resulted R2 is plot to track the convergence of the model.







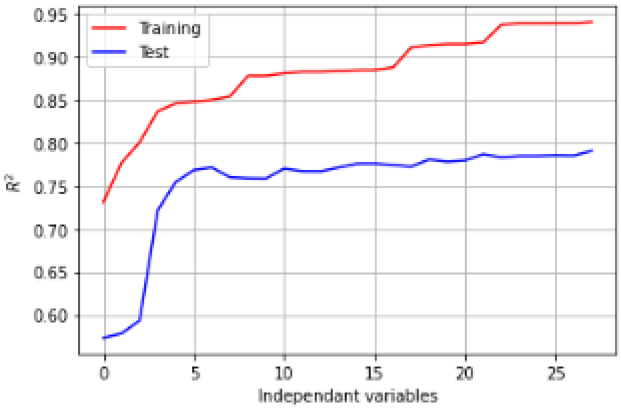
Project Description

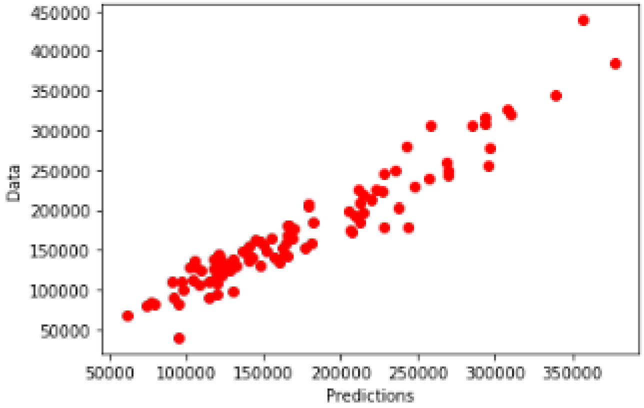
To try to improve the model made by Pr. Williams, more than 3 of the most correlated variables with SalePRice will be considered. A for loop ranging from 1 to 40 will be introduce to iterate on the number of variable included to train the linear regression model. Then the R2 will be plot to track the convergence of the model.

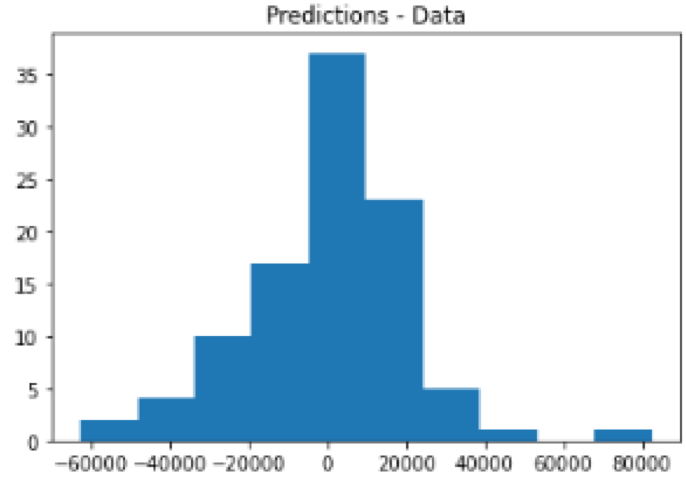
*Please add slides as required.

Verification

It can be seen that the model performs less than expected on the test data set. After 8 variables it starts to plateau.







Conclusion

To conclude, the linear regression model performs well on this data set. Including more correlated independent variables improves the predictions. However, after 8 variables, the prediction starts to plateau.

The next steps of this project are the following: compute the p-value of the linear regression, addimensionalize each variables before feeding them to the model and manipulate each variable in a way that they will all follow a normal distribution.

*Please add slides as required.