Predictive Analytics

Anosh S

TASK A – HOUSING SALE PRICES

INTRODUCTION

Predictive analytics can empower investor decision making in the real estate market with knowledge that can result in more strategic plays to maximise return. To be more specific, through investors being able to predict the price of a house based on particular characteristics, an evaluation of risks and rewards can be derived from predicted returns. Therefore, the benefits of machine learning model predictions is that it works in favour for both housing buyers and sellers.

Typically, investor evaluations of houses have required a natural intuition and manual brute force comparison of housing characteristics to understand the value of particular houses. Thus, an individual trying to enter the real estate market can be burdened and easily intimidated by the large quantities of data available, for self-research. Hence, it is the appeal of predictive analytics in being able to automate most housing research. As a result, this should ease evaluations overall through greater efficiency, and offering potentially more reliable and accurate means to determining house prices.

Overall, four different models were constructed, each producing varying results. The performance difference between these models were relatively large, but understandable when examining the underlying concepts of each of the models. Therefore, the predictive capabilities of the best model could be used to its greatest potential, and its benefits fully expounded upon, if the limitations of the model are understood and the appropriate domain knowledge is applied when drawing any conclusions.

EXPLORATORY DATA ANALYSIS

Insight into the training and test dataset, and the training set has 1570 observations while the test set has 1210 observations. Moreover, there are 80 different possible features that can be used to predict "SalePrice".

```
Train dataset has 1570 rows and 81 columns.
Test dataset has 1210 rows and 80 columns.
```

(figure 1)

Examining into whether NA values would be a problem gave that with none being present it would not be an issue of concern. This should effectively ease the analytical process with less cleaning required.

There are a number of columns with missing values. Once the best features have been selected the observations containing the NA values will be identified and dealt with accordingly.

```
There are 24 columns in train dataset with missing values.
```

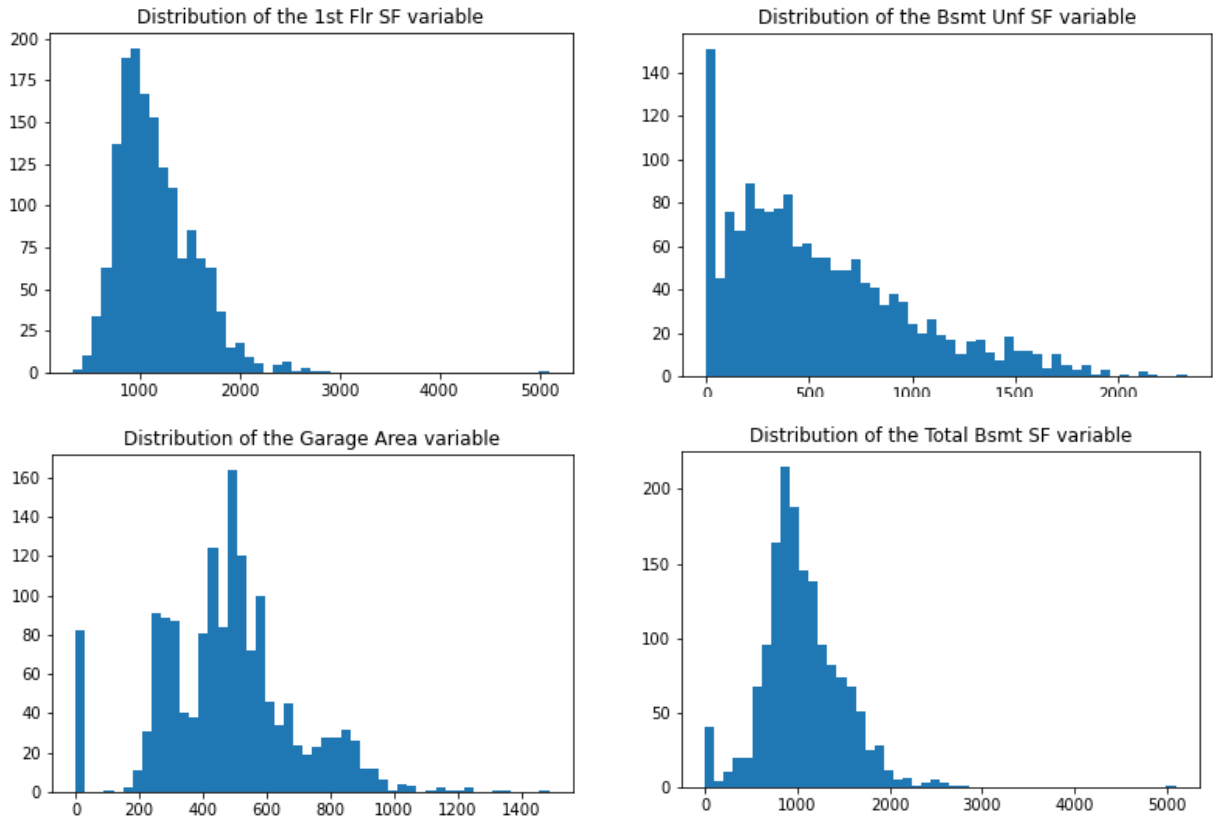
(figure 2)

Identifying the datatypes of each column, there are a number of qualitative variables that may require dummy variables to be created for these. This will be applied after an individual inspection of each of the variables in feature engineering.

```
MS SubClass           int64
MS Zoning            object
Lot Frontage        float64
Lot Area              int64
Street               object
Alley                object
...
```

(figure 3 - view full output in appendix i)

Before identifying the most ideal features, we will first delve into each of the variables to ensure that the data is logically consistent i.e. the values correspond logically with the variable that the data is trying to describe.



(figure 4 – refer to all other plots in appendix ii)

Nothing too unusual is noticeable in the data, however there is some skewness to a number of the features.

Using an IQR approach to identify outliers, these should be noted.

```
The dataset has 966 outliers.
```

(figure 5)

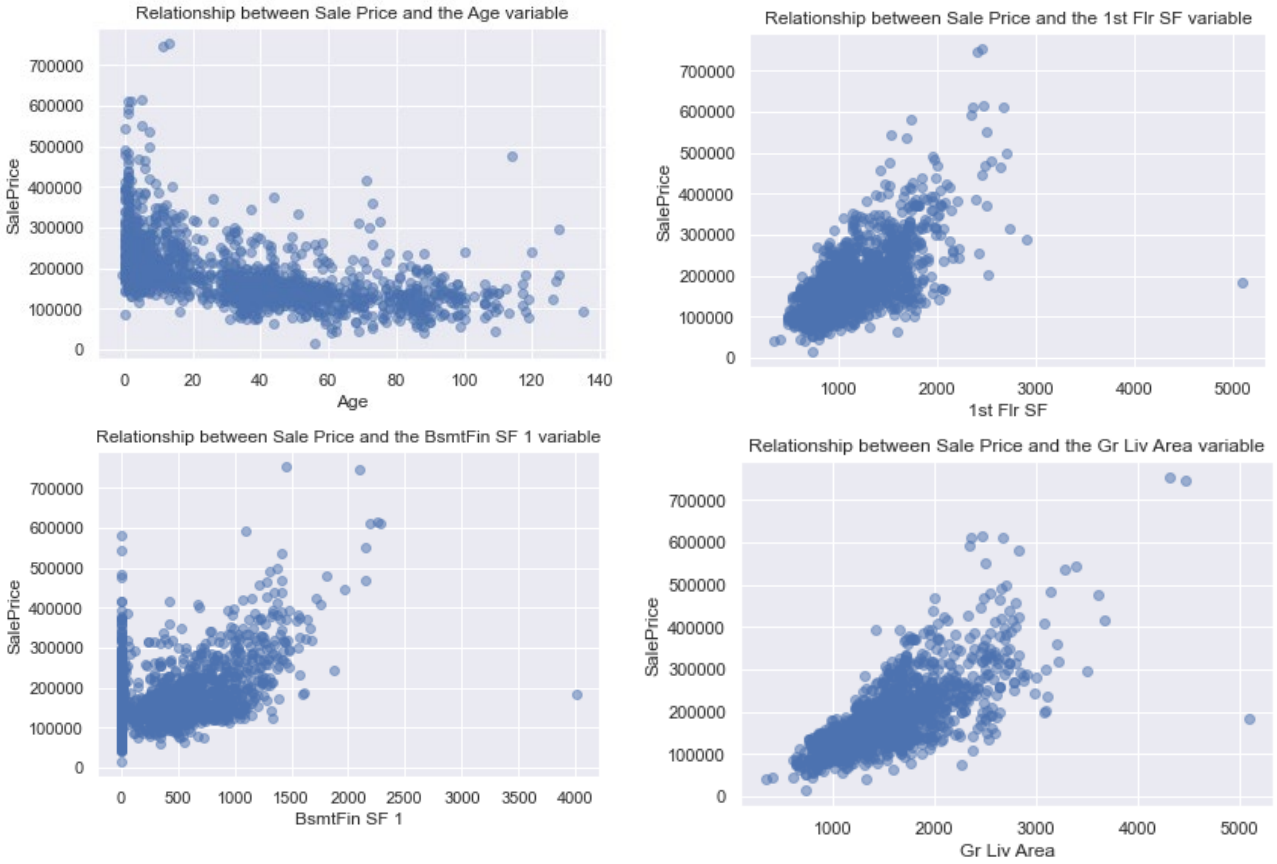
Since, there were 966 observations that contain at least one outlier under one of its features, this represents a significant proportion of the train dataset and thus were not ignored in training.

FEATURE ENGINEERING

First, we will examine for the most relevant numerical variables in the dataset. Later, the best selected categorical features will be converted to dummies so that they can be usable for predictive modelling.

Subtracting the year it was sold in to the year it was constructed the new variable 'Age' was created. Also, subtracting the last year it was remodeled or had additions made to the house to the year it was sold was added as the "Last remod/add" feature. The same was also done for "Garage Yr Blt".

The following is a series of scatter plots of the numerical features against "SalePrice" to understand what relationships may be present between these variables.



(figure 6 – refer to all other plots in appendix iii)

Viewing the scatter plots the best variables that were considered were "Gr Liv Area", "Garage Area", "Age", "Last remod/add", "Total Bsmt SF", "1st Flr SF", '2nd Flr SF", “BsmtFin SF 1”, “Full Bath”, “TotRms AbvGrd”, “Fireplaces” and “Garage Cars”. This is due to these showing some form of a non-linear relationship.

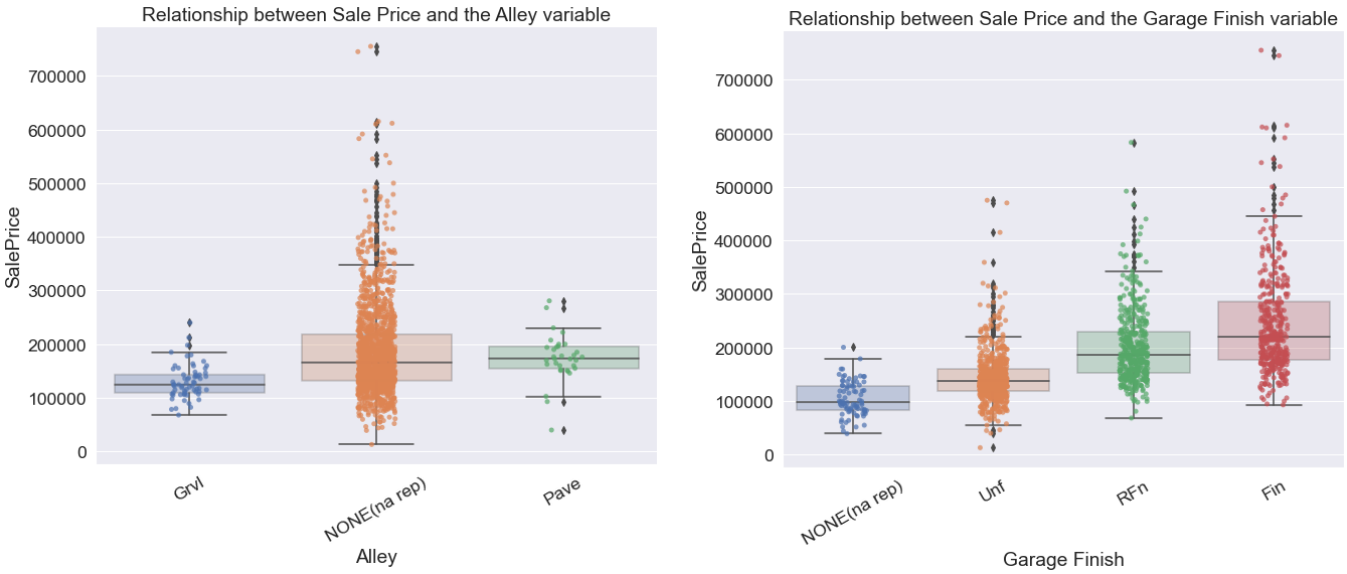
Since the scatter plots of “Age” and “Garage Age” appeared to have very similar scatter plots these variables were compared for multicollinearity to ensure linear models were not impacted in model construction.

```
0.8281
```

(figure 7)

Given that the correlation between the two was found to be relatively high the potential for these to cause potential multicollinearity issues for linear models meant these were removed for linear model feature subsets.

To see if any relationship is present between the categorical variabels and the “SalesPrice” predictor, box plots were made between these.



(figure 7 – refer to all other plots in appendix iv)

The best categorical features would be "Kitchen Qual", "Garage Finish" and "Overall Quality". This was on the basis that there was a sufficient number of observations between the different categories, a clear pattern was evident and that the variation (the whiskers of the boxplots) between groups did not overlap excessively. "Alley" would have also been considered, however, the presence of NA values that did not show a pattern or a distinct variation between the other categories meant NA values had little meaning.

Before confirming the final subset of features, to ensure that there were no issues in training and evaluation the features that contained missing values were to be dropped or imputed with an average.

Provided that a small proportion of observations had an NA value for "Garage Age", "Garage Area" and "Garage Cars" it was best to impute the values of these rows with their means rather

than removing the variables. Moreover, provided that only one observation in the test set had an NA value for "Garage Area" and "Garage Cars" it was best to impute the values of this single row with their means rather than removing the entire "Garage Area" and “Garage Cars” variables. Also, it was noticed that during model evaluation some of the observations in the test dataset under "Kitchen Qual" contained observations that had a category not defined in the training set. Hence, “Kitchen Qual” was disregarded from both sets.

Therefore, the following features that were considered for model construction and evaluation were "SalePrice", "Gr Liv Area", "Garage Area", "Age", "Last remod/add", "Total Bsmt SF", "1st Flr SF", "Kitchen Qual", "Overall Qual", "Garage Finish", '2nd Flr SF', 'BsmtFin SF 1', 'Full Bath', 'TotRms AbvGrd', 'Fireplaces' and 'Garage Cars'.

To finish feature engineering the train and test narrowed dataframes were finalised with its numerical variables being standardised. Typically the categorical variables would have been dummified here too, however, to simplify the model construction workflow this was done just before the subset of features were used to train the model.

Methodology and Modelling

kNN Regression

As the third best model and one that was studied during the unit, the kNN regression method achieved reasonable results. It is a non-parametric supervised learning approach that requires no distribution assumptions. The model predicts for some input point x as:

$$\hat{f}(x) = \frac{1}{k} \sum_{x_i \in N_k(x,D)} y_i$$

where D is the training sample i.e. $D = \{(y_i, x_i)\}_{i=1}^{N}$. Therefore, the model predicts based on the sample average of the response values for the k number of training observations that are closest to the point x.

Using the kNN model formulation best subset selection was done to determine the best model from the best features derived from feature engineering. However, given that conducting best subset selection across fifteen features is highly time consuming, only different combinations of the features were trained for models containing twelve, thirteen, fourteen or fifteen features. Also for each of the different feature subsets, these were modelled against different values of k, 1 to 50, to find the most optimal k across each of the models.

```
-----------------------------------------
-----------------------------------------
------OVERALL BEST MODEL BY CV-RMSE--------
-----------------------------------------
-----------------------------------------
Overall Best kNN model uses the features ['Gr Liv Area', 'Garage Area', 'Age',
'Last remod/add', 'Total Bsmt SF', '1st Flr SF', 'Garage Finish', '2nd Flr SF',
'Full Bath', 'Fireplaces', 'Garage Cars', 'Garage Age']
CV RMSE: 32680.7791
Number of neighbours: 6
```

(figure 8)

After formulating each of the different model possibilities the best model used features 'Gr Liv Area', 'Garage Area', 'Age', 'Last remod/add', 'Total Bsmt SF', '1st Flr SF', 'Garage Finish', '2nd Flr SF', 'Full Bath', 'Fireplaces', 'Garage Cars' and 'Garage Age' with the k set to 6. This was decided as the most optimal, since it achieved the lowest cross-fold validation RMSE of $32680.7791.

Since, the kNN model is non-parametric there are no assumptions made on the underlying data distribution, and so it is possible to accept the above model as a valid alternative means to predict housing sale prices.

Random Forest Regression

As an extension of decision tree regression, this formulation is able to improve markedly on performance through removing much of the over fitting that comes from fitting a single tree. Essentially, it is a robust supervised learning method that uses averaging across a number of randomly fitted decision trees to make a final prediction. Therefore, given a B number of trees in a random forest, to predict some point x we take the average of the predictions made by the B number of trees,

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

where $\hat{f}_{rf}^{B}(x)$ is the final prediction of the random forest and $T_b(x)$ is the prediction given by the b^{th} tree of the random forest. (Random Forest and Its Implementation, 2020)

In a random forest trees are formed from sub-samples of the data that are derived from bootstrapping aggregation techniques. This allows for reduced variance whilst keeping bias the same, optimising generalisation ability (2020).

Utilising the same best subset selection methods in kNN regression, random forest regression also trained various models based on different combination feature subsets of the best fifteen variables. However, like before this was only of subsets containing twelve, thirteen, fourteen or fifteen features.

```
-----------------------------------------
------OVERALL BEST MODEL BY CV-RMSE--------
-----------------------------------------
Overall Best RF model uses the features ['Gr Liv Area', 'Age', 'Last remod/add
', 'Total Bsmt SF', '1st Flr SF', '2nd Flr SF', 'BsmtFin SF 1', 'Full Bath', 'T
otRms AbvGrd', 'Fireplaces', 'Garage Cars', 'Garage Age']
CV RMSE: 31138.1507
-----------------------------------------
-------SECOND BEST MODEL BY CV-RMSE--------
-----------------------------------------
Second Best RF model uses the features ['Gr Liv Area', 'Garage Area', 'Age', 'L
ast remod/add', 'Total Bsmt SF', '2nd Flr SF', 'BsmtFin SF 1', 'Full Bath', 'To
tRms AbvGrd', 'Fireplaces', 'Garage Cars', 'Garage Age']
CV RMSE: 31244.5667
```

(figure 9)

The best two models were formulated with random forest regression. The lowest attaining CV RMSE was $31138.1507 using the features 'Gr Liv Area', 'Age', 'Last remod/add', 'Total Bsmt SF', '1st Flr SF', '2nd Flr SF', 'BsmtFin SF 1', 'Full Bath', 'TotRms AbvGrd', 'Fireplaces',

'Garage Cars' and 'Garage Age'. The next best model with random forest regression achieved a CV RMSE of $31244.5667 with the same set of features but with 'Garage Age' replaced with '1st Flr SF'.

As a non-parametric model, similar to kNN regression, random forest regression has no formal distributional assumptions. Hence, the above best two selected models are valid.

Multiple Linear Regression

Applying the ordinary least squares method, the multiple linear regression model used to explain housing sale prices was derived through selecting the coefficient values that minimise the residual sum of squares,

$$\widehat{\beta_{ols}} = argmin_{\beta} \sum_{i=1}^{p} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

This is such that we are able to estimate for a linear equation to predict some response Y:

$$Y = \beta_0 + \beta_1 X + \beta_2 X + \cdots + \beta_p X + \epsilon$$

Again, we used the same method of best subset selection as the other models. However, as discovered earlier, "Garage Age" was highly correlated with "Age", and as a means of avoiding potential multicollinearity issues the "Garage Age" variable was excluded from subset selection.

```
-----------------------------------------
------OVERALL BEST MODEL BY CV-RMSE--------
-----------------------------------------
Overall Best MLR model uses the features ['Garage Area', 'Age', 'Last remod/add
', 'Total Bsmt SF', '1st Flr SF', 'Garage Finish', '2nd Flr SF', 'BsmtFin SF 1
', 'TotRms AbvGrd', 'Fireplaces', 'Garage Cars']
CV RMSE: 38232.5707
```

(figure 10)

The best linear regression model was only able to achieve a CV RMSE of $38232.5707, which is considerably more than the other earlier discussed models. This was achieved using the features 'Garage Area', 'Age', 'Last remod/add', 'Total Bsmt SF', '1st Flr SF', 'Garage Finish', '2nd Flr SF', 'BsmtFin SF 1', 'TotRms AbvGrd', 'Fireplaces' and 'Garage Cars'.

While the above may be the best model using linear regression, there are underlying assumptions that must be satisfied in order for it to be valid for use in predictions. Therefore, to make any conclusions from the above linear model, a checking of assumptions would be first required.

Polynomial Regression

As a means to increase the complexity of the linear model, polynomial regression may be able to address underfitting of the linear model. Utilising the same linear regression and least squares principles, polynomial regression is essentially linear regression except that it includes the same features of higher power, depending on the degree of the polynomial.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^d + \epsilon$$

Similarly, to all the other modelling methodologies the same procedure in best subset selection was applied here too. Also, to derive the best polynomial degree across the different feature subsets different values of d, 1 to 3, were used.

```
-----------------------------------------
------OVERALL BEST MODEL BY CV-RMSE--------
-----------------------------------------
Overall Best POLY model uses the features ['Garage Area', 'Age', 'Garage Age',
'Last remod/add', 'Total Bsmt SF', '1st Flr SF', 'Garage Finish', '2nd Flr SF',
'BsmtFin SF 1', 'TotRms AbvGrd', 'Fireplaces', 'Garage Cars'] best degree is 3
CV RMSE: 38253.1741
```

(figure 11)

Therefore, the best polynomial model had features 'Garage Area', 'Age', 'Garage Age', 'Last remod/add', 'Total Bsmt SF', '1st Flr SF', 'Garage Finish', '2nd Flr SF', 'BsmtFin SF 1', 'TotRms AbvGrd', 'Fireplaces' and 'Garage Cars', and was of degree 3. As a result, it achieved a CV RMSE of $38253.1741.

Similarly, as a parametric model there are assumptions that must be satisfied in order for any conclusions to be made of the polynomial model. These are that the independent variables should be independent of each other, and that the errors are independent, normally distributed with mean zero and of constant variance. Once these checks have been confirmed then it would be a valid model for predictive use.

Model Evaluation

	CV RMSE	Test RMSE
RF1	31138.1507	27721.3746
RF2	31244.5667	28168.5235
kNN	32680.7791	32801.2053
MLR	38232.5707	43209.2730
PR	38253.1741	43228.9505

(figure 12)

Having performed model construction, the following delves into the ability for the each of the best models to generalise with unseen data i.e. test data. Across the best models of each of the types, the best overall was the first best random forest model attaining both a low CV RMSE and test RMSE of $27721.3746. Following this is the next best random forest model achieving the second lowest CV RMSE and also second lowest test RMSE of $28168.5235. Then in both CV RMSE and test RMSE the following are the next best in their respective order: kNN, multiple linear regression and then polynomial regression. Considering, the cross validation and test RMSEs being relatively similar, these imply the bias-variance tradeoff is being most optimally met for each of the different model types and thus also the best that has been selected of them.

The best performance being attained by the random forest models can be best explained by its use of bootstrapping aggregation techniques to build de-correlated trees. By averaging the noisy approximately unbiased decision tree models, it is possible to capture the complex interaction of patterns in the data while keeping variance low and bias the same (2020). Therefore, the random forest model most optimally generalises with a bias-variance trade off that is more optimal than kNN, linear and polynomial regression.

kNN regression comes in as third best, behind the two random forest models, as the data seems most suitably modelled by non-linear methods. Moreover, the use of local methods by kNN regression means bias is also minimised, while its additive nature keeps variance low as well (Bradsher, 2020). However, the bias-variance tradeoff is not as optimal as random forest regression, since kNN performs poorly with outlier and noisy data, while random forests are more robust to both (Kumar, Kumar and profile, 2020).

The linear and polynomial regression models perform similarly but worse than the other two types. Linear regression obviously does worse mainly in part of its failure to identify non-linear patterns. However, in an attempt to resolve this using polynomial regression performance is still slightly worse than linear regression in both cross validation and test RMSE. This could be due to the polynomial degree not being high enough to capture enough of the complexities that the linear model fails to account for. Moreover, there is the additional consideration of assumptions that come with polynomial and linear regression models that are possibly undermining the overall effectiveness of these models.

Therefore, the random forest models are the best performers overall, since it is the most robust to outliers and noisy data and is model that is able to capture non-linear complexities while maintaining a suitable bias-variance tradeoff. Unlike random forest regression, kNN is not robust to outlier and noisy data, and polynomial and linear regression models are not sufficiently complex.

Conclusion

In conclusion, different model types gave varying results and thus interesting insights into the shortfalls of particular model formulations in predicting house sale prices. While the predictive capabilities of the random forest show the most promise in practical use, there are still limitations of the model that must be noted. Improvements to consider is that data is more complete, where although only a small proportion of the dataset required the imputing of mean values these may have impacted performance. In addition to having more complete data, the "Kitchen Qual" feature being dropped due to training and test set inconsistencies meant a huge loss of useful data. Another aspect that was not delved into was the adjustment of hyperparameters concerning the random forest model, these include max depth, the number of trees, minimum sample split, etc. Thus, the lacking experimentation of this aspect means the possible overlooking of potentially better models. Also, while the features selected for the best model are the best predictors to explain the housing sale prices, it must be acknowledged that these do not explain a causal relationship with housing sale prices but rather only an association. Therefore, from the limitations the most important considerations for future research is the need to experiment with hyperparameters of the random forest model and the need to have more complete data or better value imputation through possibly using regression to predict those missing values to possibly derive more improved results in predicting for housing sale prices.

Hence, while these models are definitely useful tools, these should not be the sole reasons to any final decisions. This requires that the users of these tools makes use of domain knowledge and takes into account other factors that may not be considered by them, such as the constantly changing circumstances of sale prices in the housing market.

TASK B – TIMESERIES FORECASTING

EXPLORATORY DATA ANALYSIS

The column and row dimensions of the dataset are as follows.

```
The dataset has 312 rows and 2 columns.
```

(figure 13)

Determining the number of NA values, none were present.

```
There are 0 columns in train dataset with missing values.
```

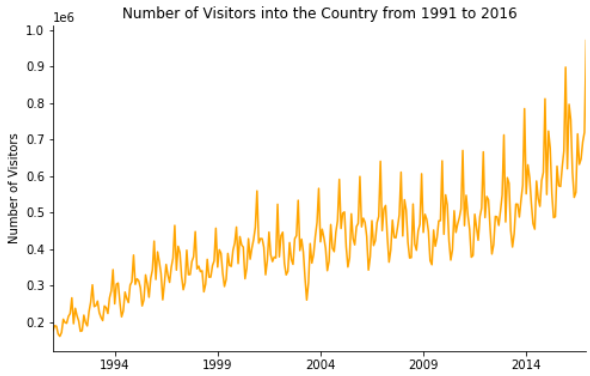
(figure 14)

The data frame was adjusted to have the date as the index column. An inspection of the head is shown below

Date	Number of Visitors
2016-08	632600
2016-09	647200
2016-10	694300
2016-11	720500
2016-12	971800

(figure 15)

The following plots the pattern change in the number of visitors entering the country from 1991 to 2016. Using this we were able to delineate the overall seasonality and trend of the data.



(figure 16)

Therefore, it was apparent that the data seemed to follow some rough yearly seasonality. A consistent upward trend was also observed while variance appeared to increase towards 2016. To understand the variations further a timeseries decomposition can be viewed as below.

```
count       312.0000
mean     419407.3718
std      132443.0593
min      161400.0000
max      971800.0000
```
(figure 17 - refer to full results in appendix v)

Moreover, the mean entrance of visitors is 419307 across the 312 months (1991-2016). Also the highest recorded number was 971800 while the lowest was 161400.

Multiplicative Holt-Winters Exponential Smoothing

Rationale

The Holt-Winters exponential smoothing model considers trend and seasonal correctional methods. Therefore, the following model was selected on the rationale of seasonality and trend variations being key patterns in the data, as observed in the EDA.

Also, it is known that the additive model is most appropriate for data that has seasonal variations that are roughly constant across the provided period, while a multiplicative one suits data when the seasonal variation is proportional to the trend. Since the EDA showed that the variance of seasonal periods increased towards 2016 it made sense to proceed with a model formulation that applied the multiplicative method.

Methodology

The selected model compromises of three different components, these being the level, trend and seasonal indices.

$$\hat{y}_{t+1} = (l_t + b_t) \times S_{t+1-L} \ (forecast\ equation)$$

$$l_t = \alpha\left(\frac{y_t}{S_{t-L}}\right) + (1-\alpha)(l_{t-1} + b_{t-1}) \ (level)$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \ (trend)$$

$$S_t = \delta\left(\frac{y_t}{l_t}\right) + (1-\delta)S_{t-L} \ (seasonal\ indices)$$

Where L is the seasonal frequency, thus, predictions of the future or of current instances are computed on the basis of these three influences.

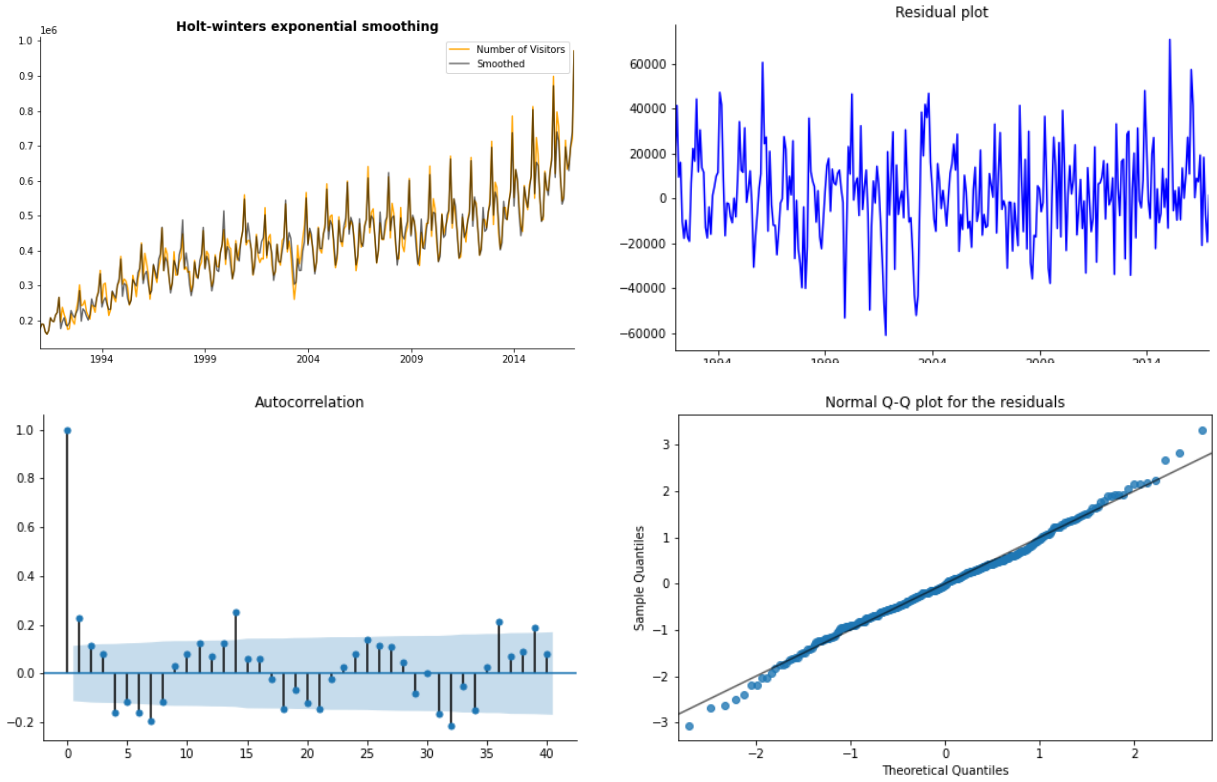
Using exponential smoothing across these different aspects that the model considers, it determines a weight for past observations. This means that the model gives greater weight to observations recorded more recent, than those recorded before them. Hence, parameters alpha, beta and delta influence the level of exponential smoothing applied to level, trend and seasonal components respectively. These values are optimised using least squares.

$$\hat{\alpha}, \hat{\beta}, \hat{\delta} = argmin_{\alpha,\beta,\delta} \sum_{t=1}^{N} (y_t - (l_{t-1} + b_{t-1}) \times S_{t-L})^2$$

Diagnostics

Before deriving any forecasts, diagnostics were performed to ensure that the model was appropriate in forecasting of the provided data.

Holt-Winters exponential smoothing curve following very closely with the data was a good indication of model use for the given data.



(figure 18)

The residuals appeared not to follow any particular pattern, however, a qqplot of them showed some slight deviation from the line, through which did raise some concern for normality. Moreover, the autocorrelations did not give any unusual results and were reasonable. Thus, provided these diagnostics, other than the slight concern for normality, the model was confirmed appropriate for the data.

Before completing model validation and selection, the next model of comparison will be discussed first.

Multiplicative Holt-Winters Exponential Smoothing with Dampening

Rationale

Since the multiplicative Holt-Winters Exponential Smoothing with Dampening model follows very similar principles to the previous model, the same rationale applies. However, the addition of dampening also allows us to address the potential issue of extrapolating the trend

indefinitely into the future. Therefore, producing such a model should be able to prevent the resulting of implausible forecasts.

Methodology

The model follows the same principles and components of the previous one, but with the inclusion of a dampening parameter (φ). Effectively, this applies an additional weight to both the seasonality and trend components.

$$\hat{y}_{t+1} = (l_t + \varphi b_t) \times S_{t+1-L} \; (forecast\ equation)$$

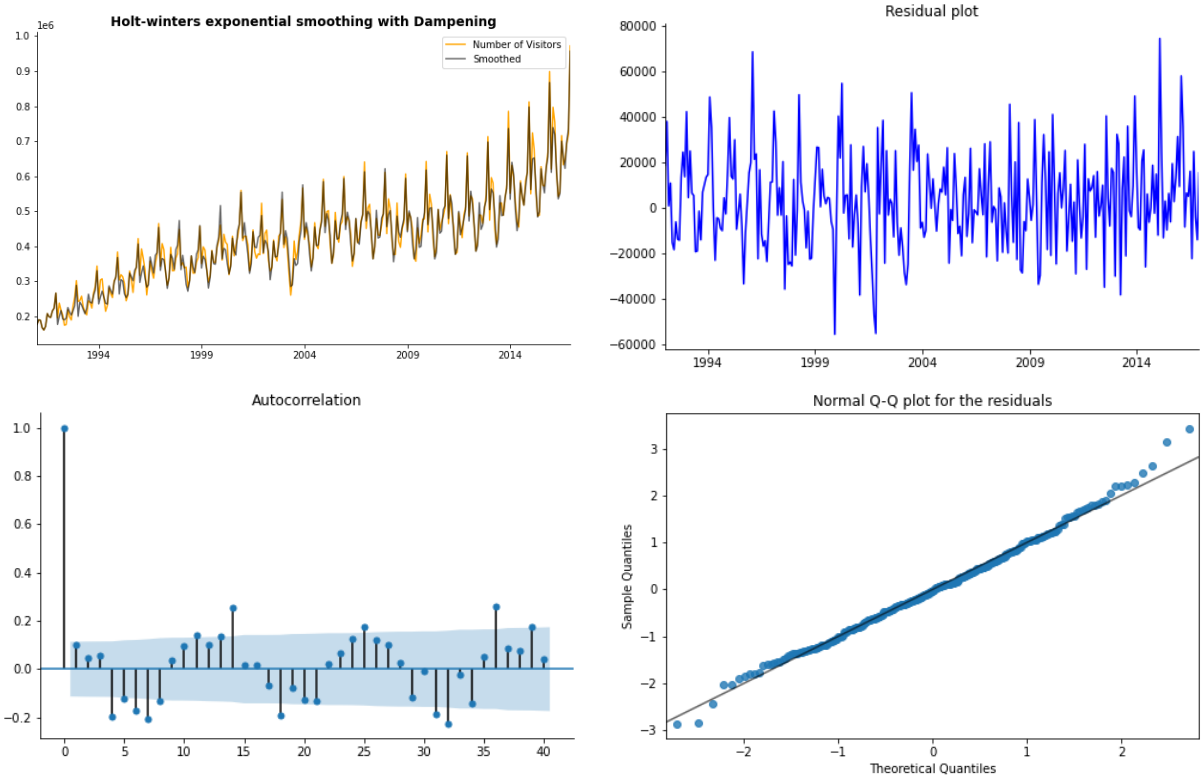
$$l_t = \alpha\left(\frac{y_t}{S_{t-L}}\right) + (1-\alpha)(l_{t-1} + \varphi b_{t-1}) \; (level)$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)\varphi b_{t-1} \; (trend)$$

$$S_t = \delta\left(\frac{y_t}{l_t}\right) + (1-\delta)S_{t-L} \; (seasonal\ indices)$$

Diagnostics

Holt-Winters exponential smoothing curve following very closely with the data was a good indication of model use for the given data.



(figure 19)

Similarly, to the previous model the residuals did not follow any patterns, while normality was also questionable but relatively reasonable with the slight deviations from the line at the

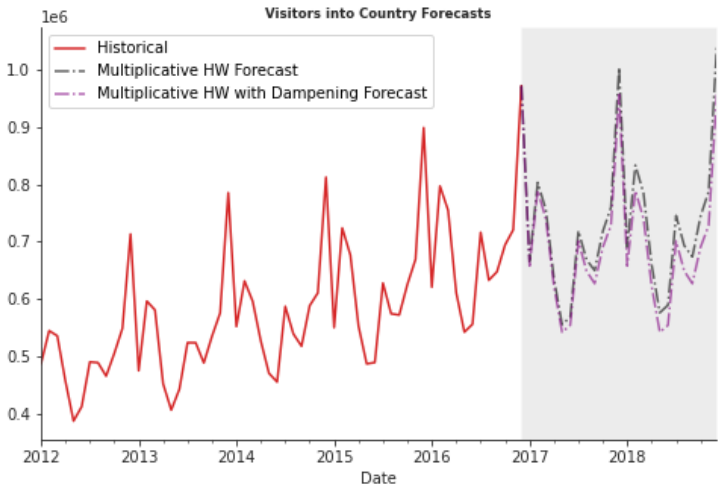
extremities. Moreover, the autocorrelations were also not an issue with these not containing any high values.

Model Validation and Selection

	RMSE	SE
Holt-Winters Multiplicative	23645.48	2140.26
Holt-Winters Multiplicative (with dampening)	24164.86	2170.61

(figure 20)

Given the results of the above models there was a slight difference in the RMSE and SE scores. The predictions were also slightly different, where the model with dampening had lower values compared to without dampening. While the differences in performance and forecasting were basically negligible the best model to consider would be the one with dampening. This is because it assesses uncertainty of the future better through not erroneously assuming indefinite growth.



Conclusion

In conclusion, the forecasting results were successfully able to forecast 24 months into future on the number of visitors that would enter the country. Although, the multiplicative Holt-Winters modelling with dampening prevents the erroneous assumption that growth continues indefinitely, the model is still limited by historical data. This means that it is not able to account for influences that go beyond historical data, such as the enforcing of new regulations or the influence of economic factors. Therefore, an improvement to consider in further research would be how this understanding of external influences can be incorporated into time series forecasting.

Hence, while forecasting models can be a useful tools, any conclusions drawn from these models must be supported with strong domain knowledge and insight of external factors that such models are limited by.

References

Bradsher, M., 2020. Why Would Anyone Use KNN For Regression?. [online] Cross Validated. Available at: <https://stats.stackexchange.com/questions/104255/why-would-anyone-use-knn-for-regression> [Accessed 19 November 2020].

Kumar, N., Kumar, N. and profile, V., 2020. Advantages And Disadvantages Of KNN Algorithm In Machine Learning. [online] Theprofessionalspoint.blogspot.com. Available at: <http://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of-knn.html> [Accessed 18 November 2020].

Kumar, N., Kumar, N. and profile, V., 2020. Advantages And Disadvantages Of KNN Algorithm In Machine Learning. [online] Theprofessionalspoint.blogspot.com. Available at: <http://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of-knn.html> [Accessed 19 November 2020].

Math.mcgill.ca. 2020. [online] Available at: <https://www.math.mcgill.ca/yyang/resources/doc/randomforest.pdf> [Accessed 18 November 2020].

Medium. 2020. Random Forest And Its Implementation. [online] Available at: <https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f> [Accessed 18 November 2020].

Scikit-learn.org. 2020. 3.2.4.3.2. Sklearn.Ensemble.Randomforestregressor — Scikit-Learn 0.23.2 Documentation. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html> [Accessed 16 November 2020].

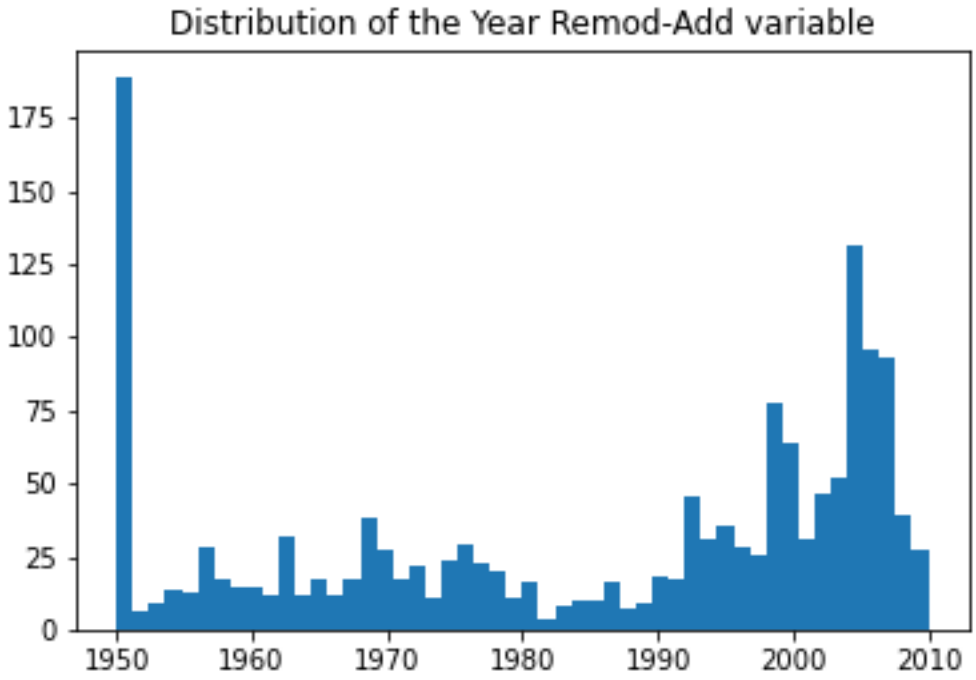
APPENDIX

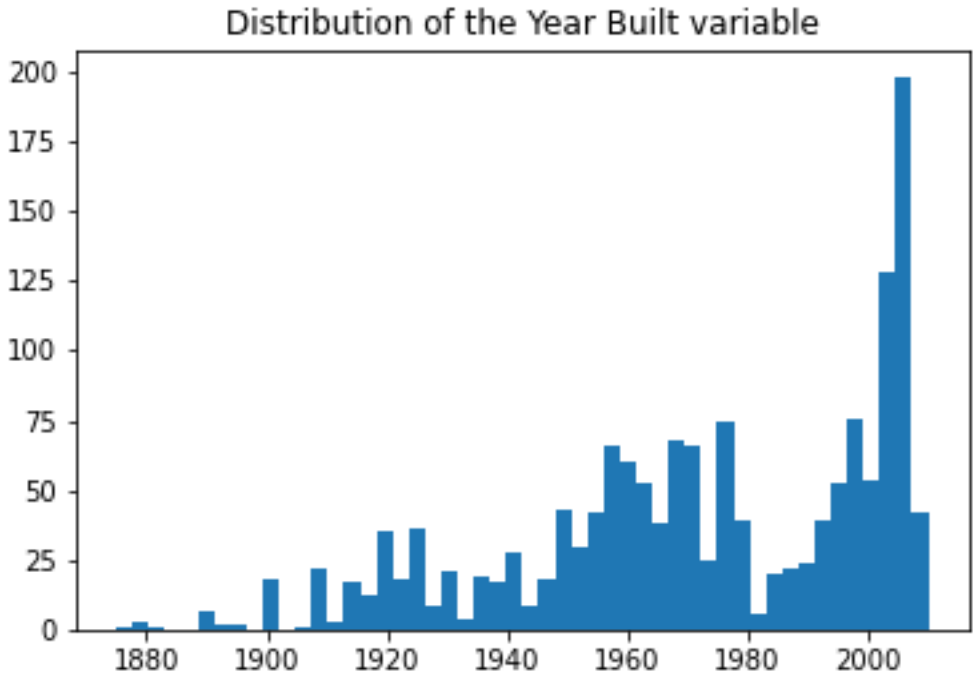
Appendix i

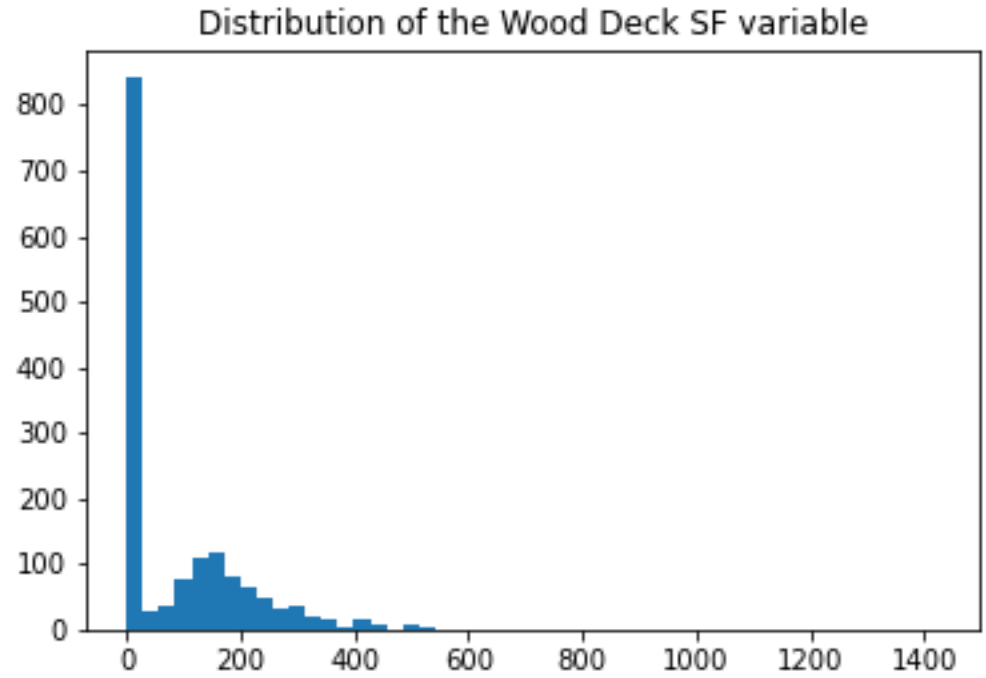
```
MS SubClass          int64
MS Zoning           object
Lot Frontage       float64
Lot Area             int64
Street              object
Alley               object
Lot Shape           object
Land Contour        object
Utilities           object
Lot Config          object
Land Slope          object
Neighborhood        object
Condition 1         object
Condition 2         object
Bldg Type           object
House Style         object
Overall Qual         int64
Overall Cond         int64
Year Built           int64
Year Remod/Add       int64
Roof Style          object
Roof Matl           object
Exterior 1st        object
Exterior 2nd        object
Mas Vnr Type        object
Mas Vnr Area       float64
Exter Qual          object
Exter Cond          object
Foundation          object
Bsmt Qual           object
Bsmt Cond           object
Bsmt Exposure       object
BsmtFin Type 1      object
BsmtFin SF 1       float64
BsmtFin Type 2      object
BsmtFin SF 2       float64
Bsmt Unf SF        float64
Total Bsmt SF      float64
Heating             object
Heating QC          object
Central Air         object
Electrical          object
1st Flr SF           int64
2nd Flr SF           int64
Low Qual Fin SF      int64
Gr Liv Area          int64
Bsmt Full Bath     float64
Bsmt Half Bath     float64
Full Bath            int64
Half Bath            int64
```

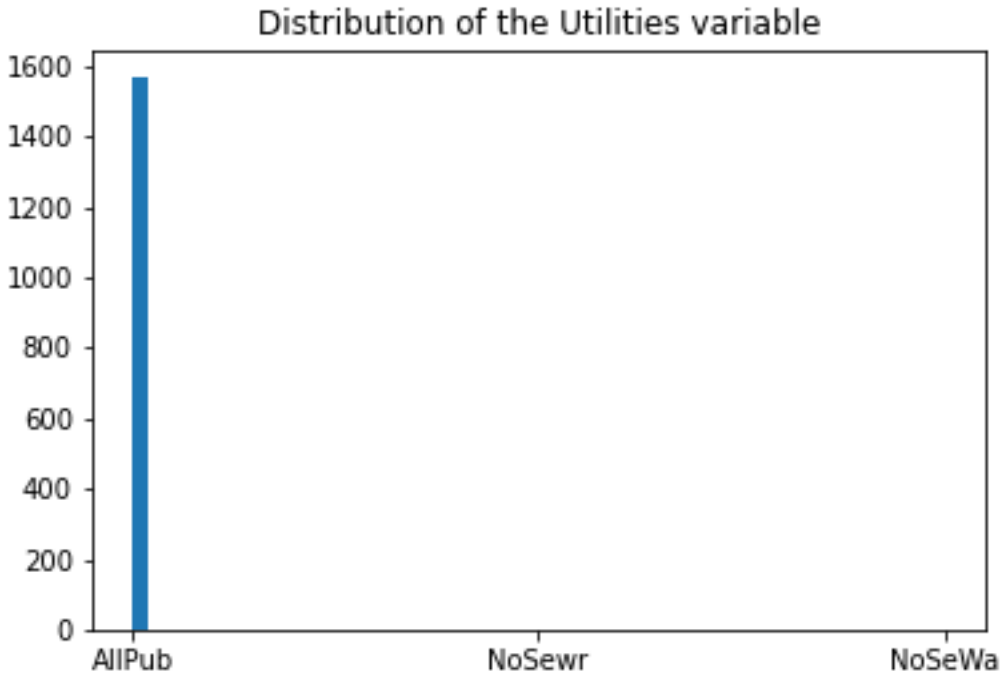
```
Bedroom AbvGr        int64
Kitchen AbvGr        int64
Kitchen Qual        object
TotRms AbvGrd        int64
Functional          object
Fireplaces           int64
Fireplace Qu        object
Garage Type         object
Garage Yr Blt      float64
Garage Finish       object
Garage Cars          int64
Garage Area          int64
Garage Qual         object
Garage Cond         object
Paved Drive         object
Wood Deck SF         int64
Open Porch SF        int64
Enclosed Porch       int64
3Ssn Porch           int64
Screen Porch         int64
Pool Area            int64
Pool QC             object
Fence               object
Misc Feature        object
Misc Val             int64
Mo Sold              int64
Yr Sold              int64
Sale Type           object
Sale Condition      object
SalePrice            int64
dtype: object
```

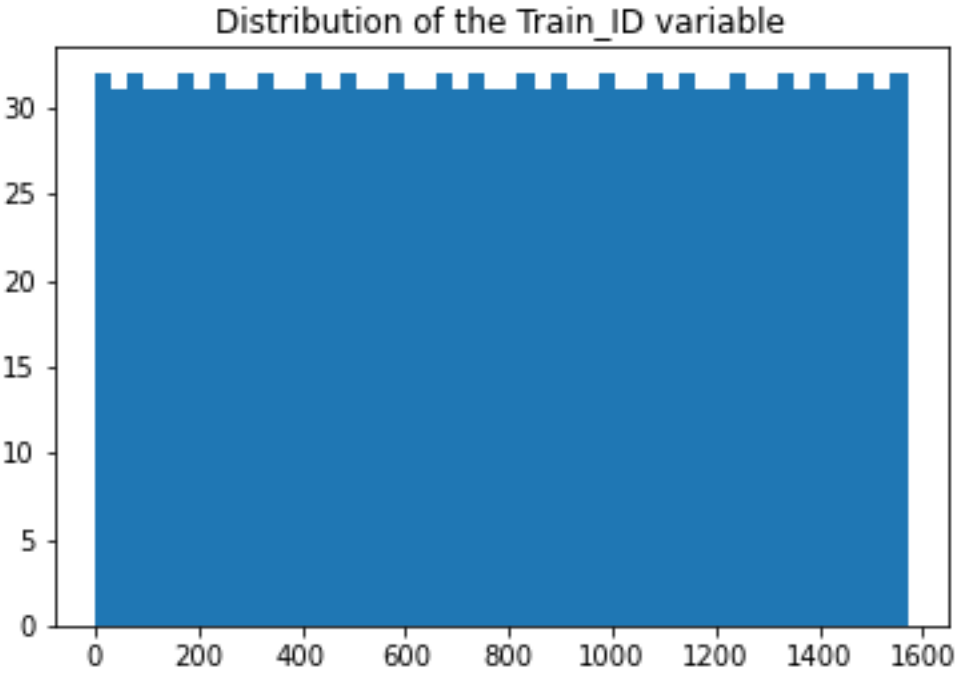
Appendix ii

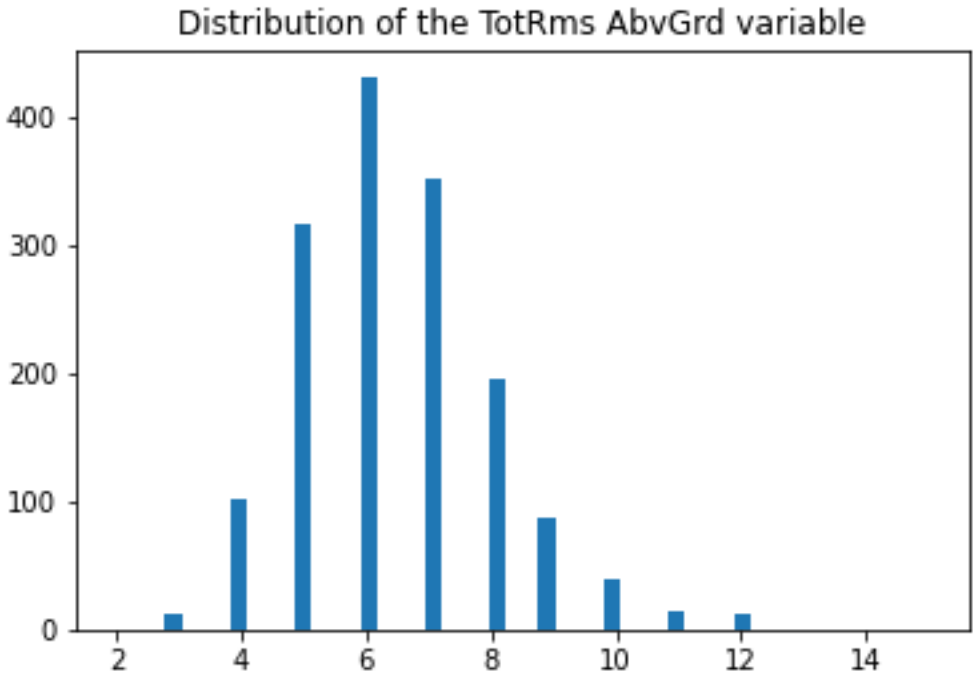

Distribution of the Year Remod-Add variable
175
150
125
100
75
50
25
0
1950
1960
1970
1980
1990
2000
2010

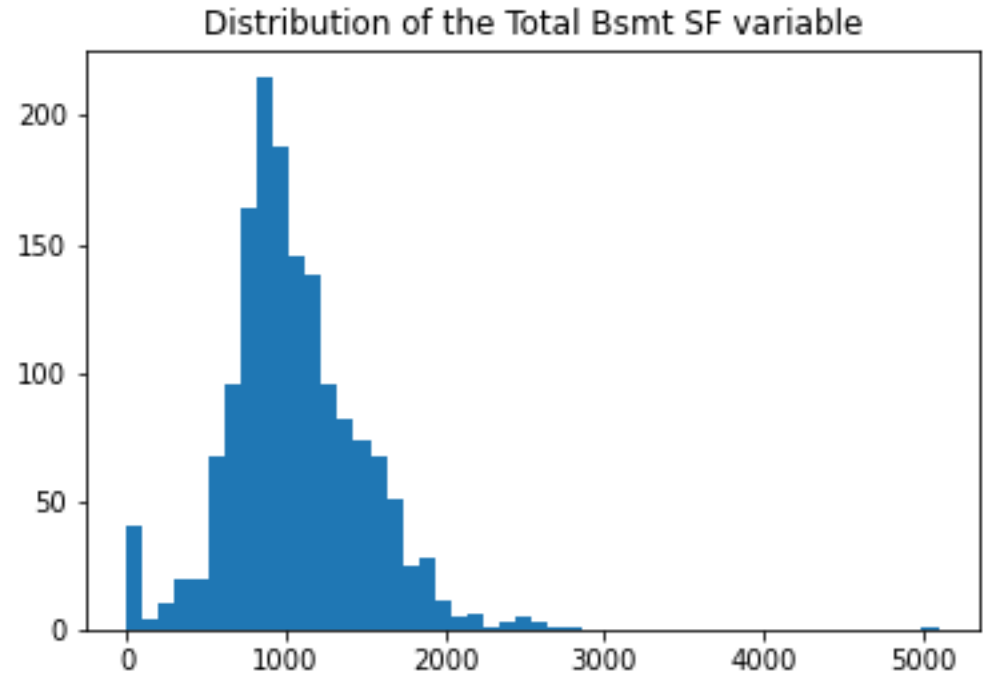

Distribution of the Year Built variable
200
175
150
125
100
75
50
25
0
1880
1900
1920
1940
1960
1980
2000

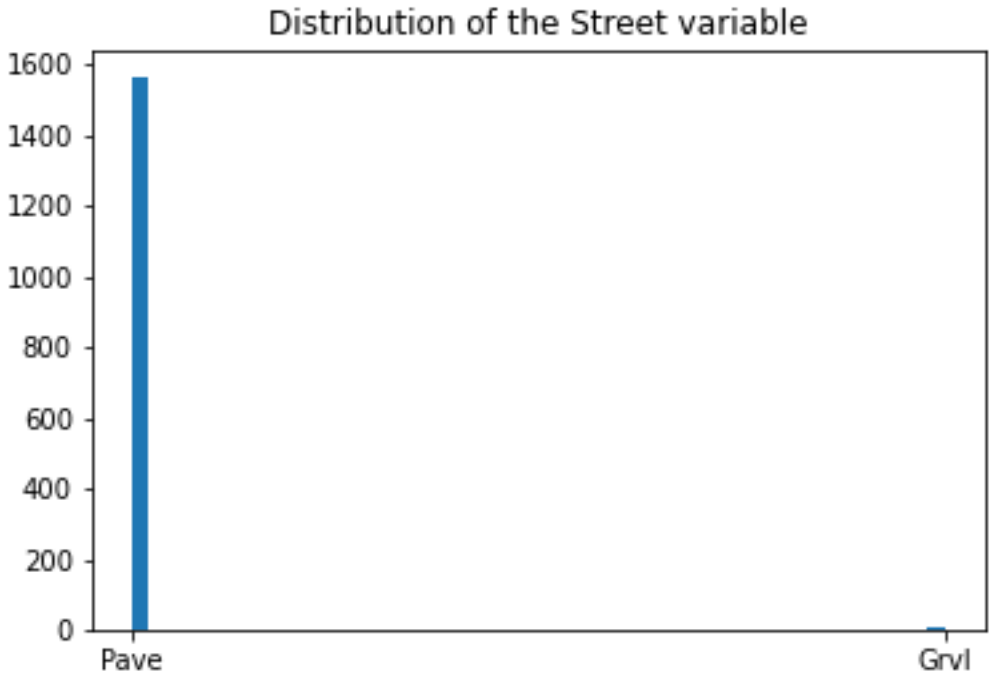

Distribution of the Wood Deck SF variable
800
700
600
500
400
300
200
100
0
0
200
400
600
800
1000
1200
1400

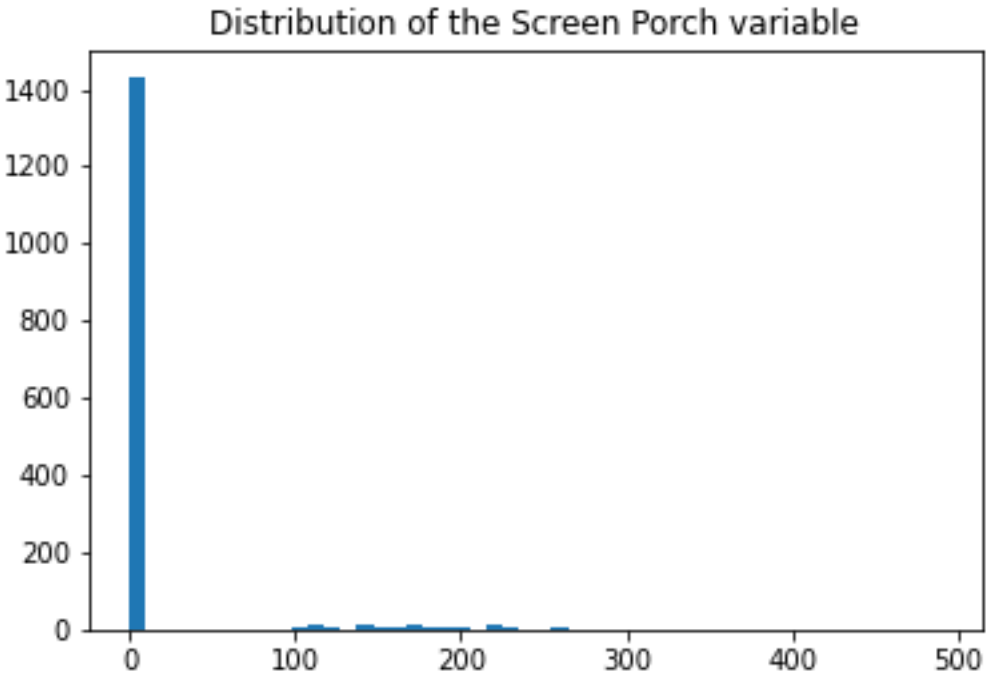

Distribution of the Utilities variable
1600
1400
1200
1000
800
600
400
200
0
AllPub
NoSewr
NoSeWa

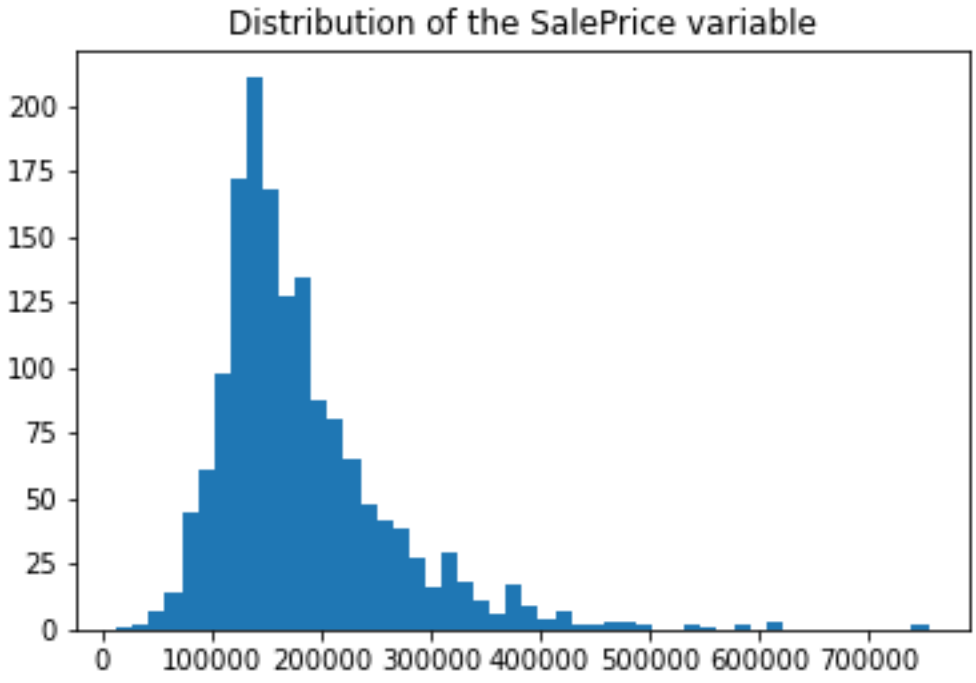

Distribution of the Train_ID variable
30
25
20
15
10
5
0
0
200
400
600
800
1000
1200
1400
1600

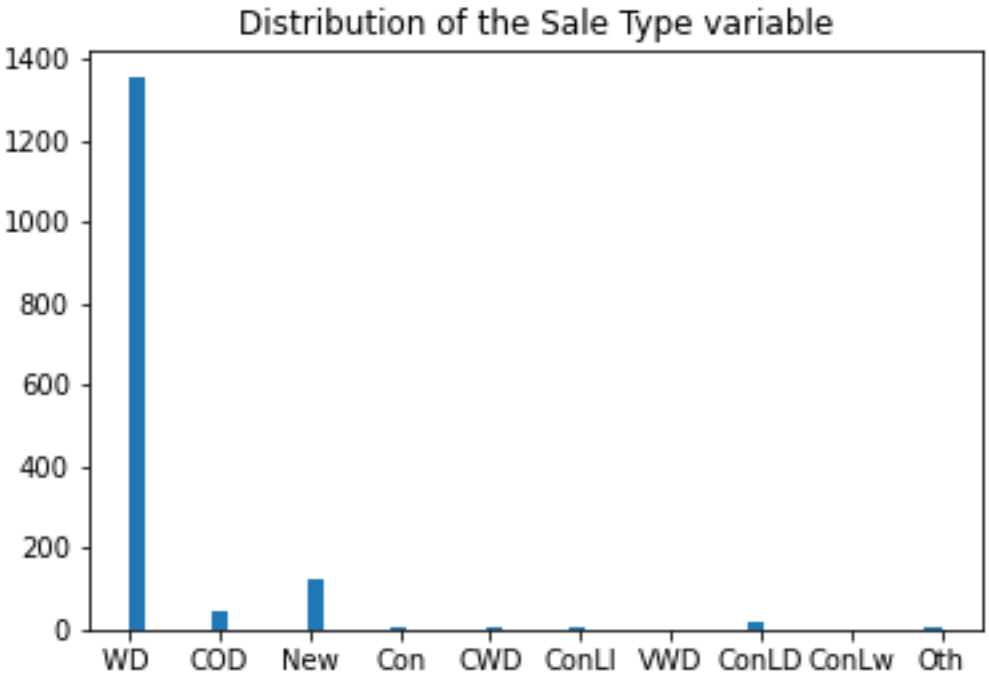

Distribution of the TotRms AbvGrd variable
400
300
200
100
0
2
4
6
8
10
12
14

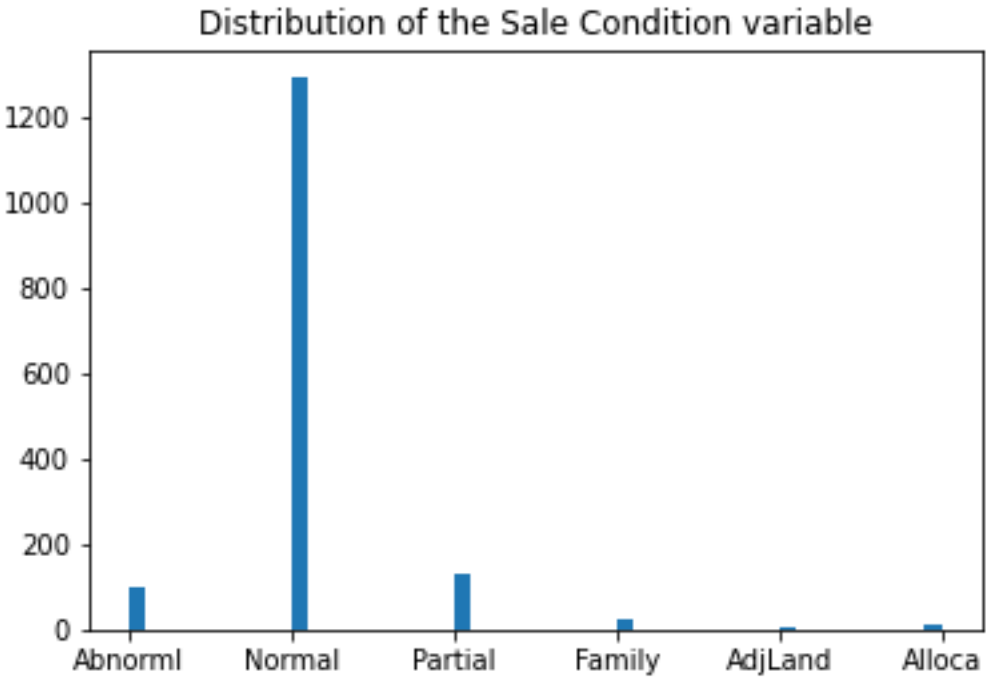

Distribution of the Total Bsmt SF variable
200
150
100
50
0
0
1000
2000
3000
4000
5000

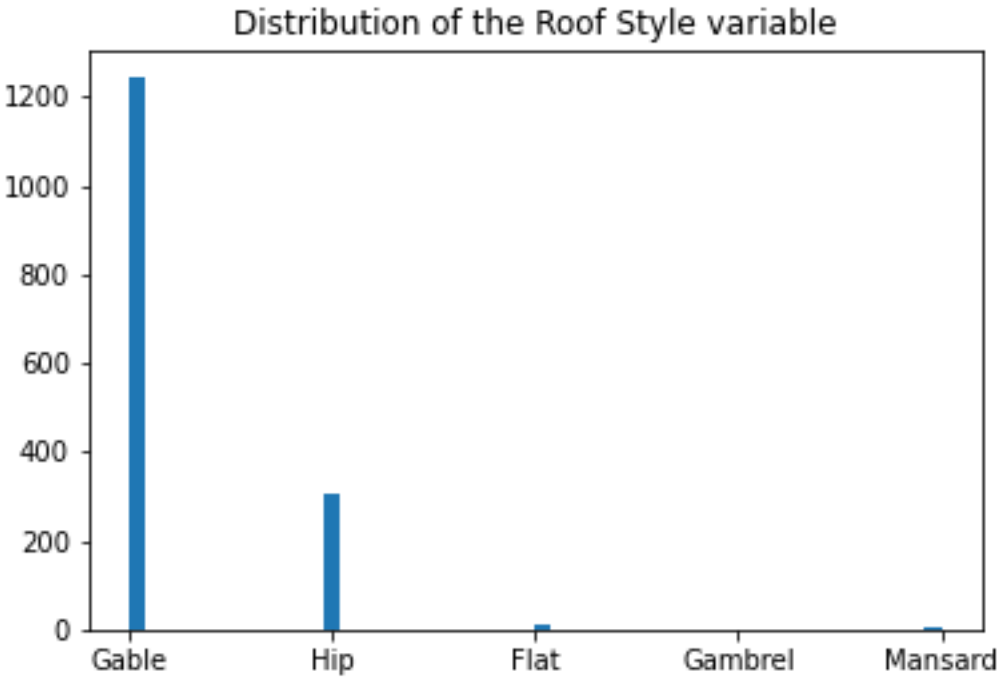

Distribution of the Street variable
1600
1400
1200
1000
800
600
400
200
0
Pave
Grvl

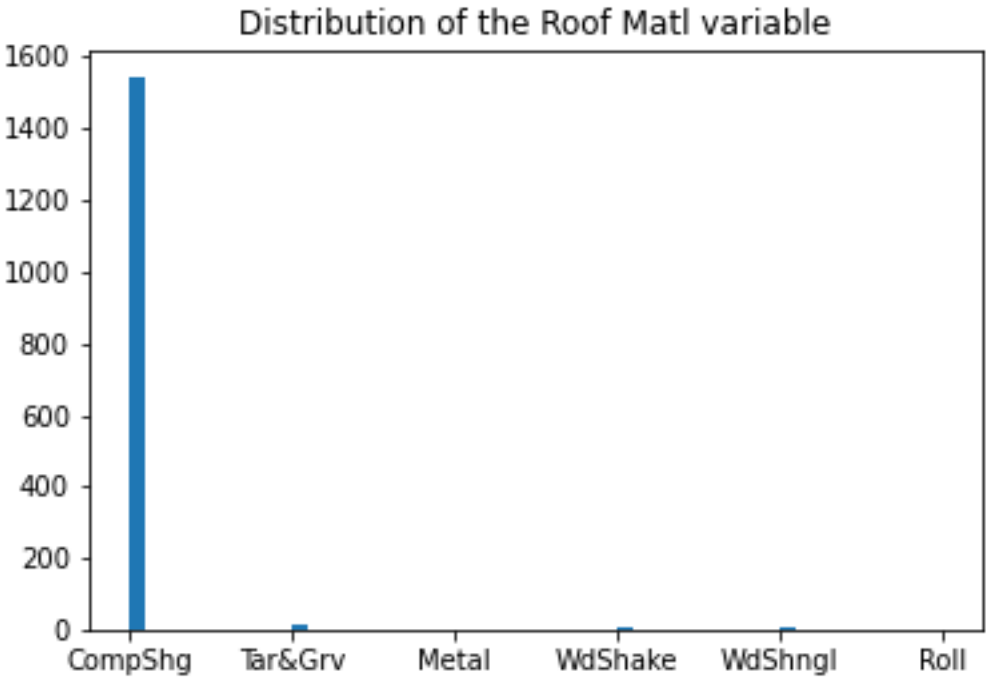

Distribution of the Screen Porch variable
1400
1200
1000
800
600
400
200
0
0
100
200
300
400
500

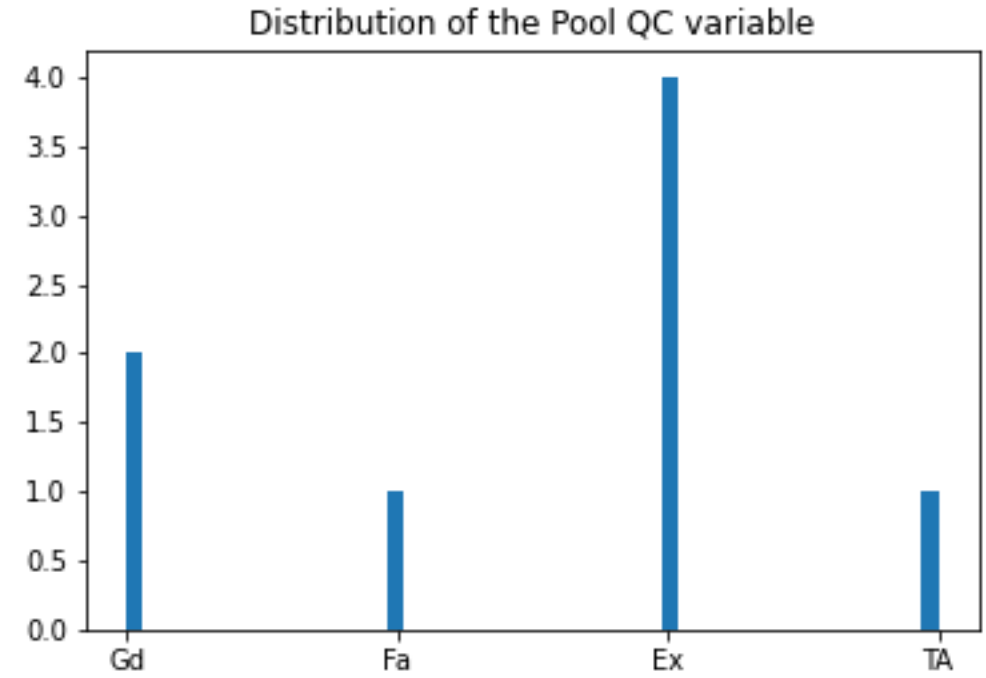

Distribution of the SalePrice variable
200
175
150
125
100
75
50
25
0
0
100000
200000
300000
400000
500000
600000
700000

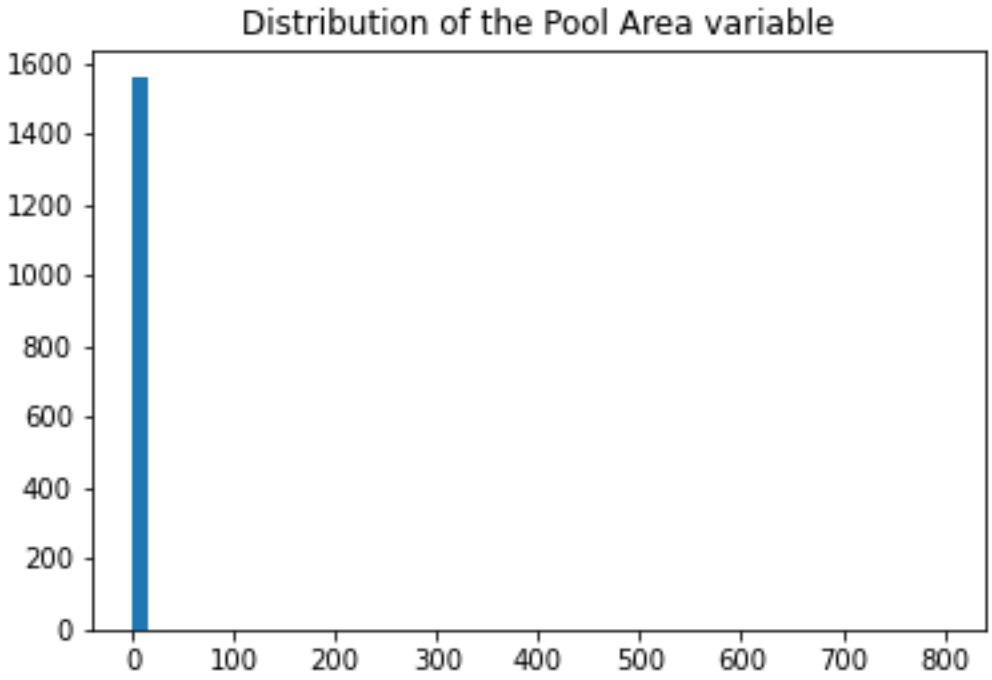

Distribution of the Sale Type variable
1400
1200
1000
800
600
400
200
0
WD
COD
New
Con
CWD
ConLI
VWD
ConLD
ConLw
Oth

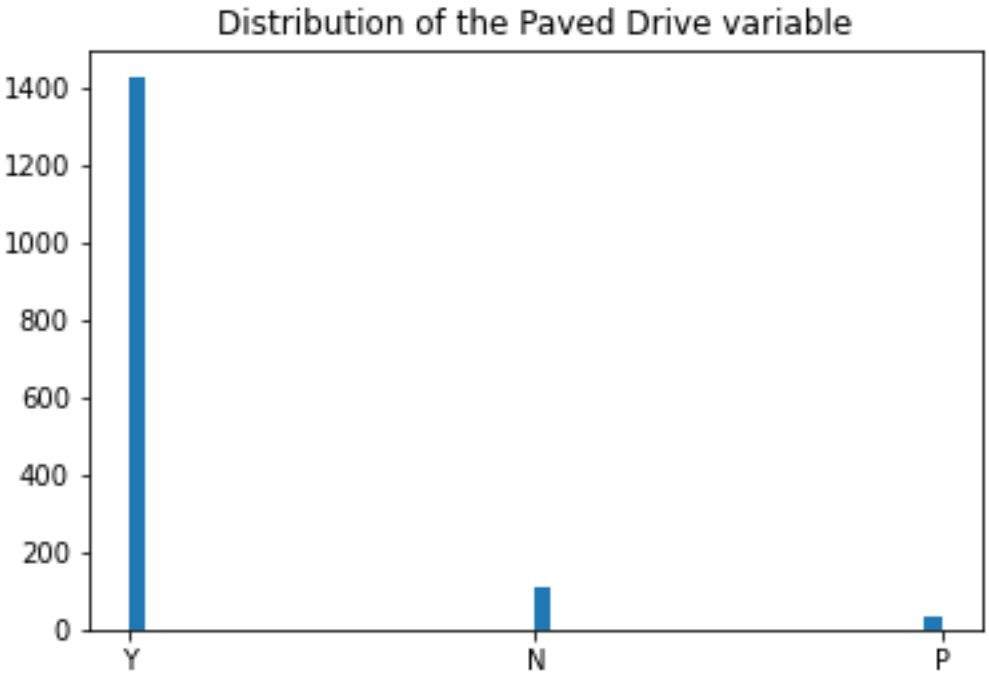

Distribution of the Sale Condition variable
1200
1000
800
600
400
200
0
Abnorml
Normal
Partial
Family
AdjLand
Alloca

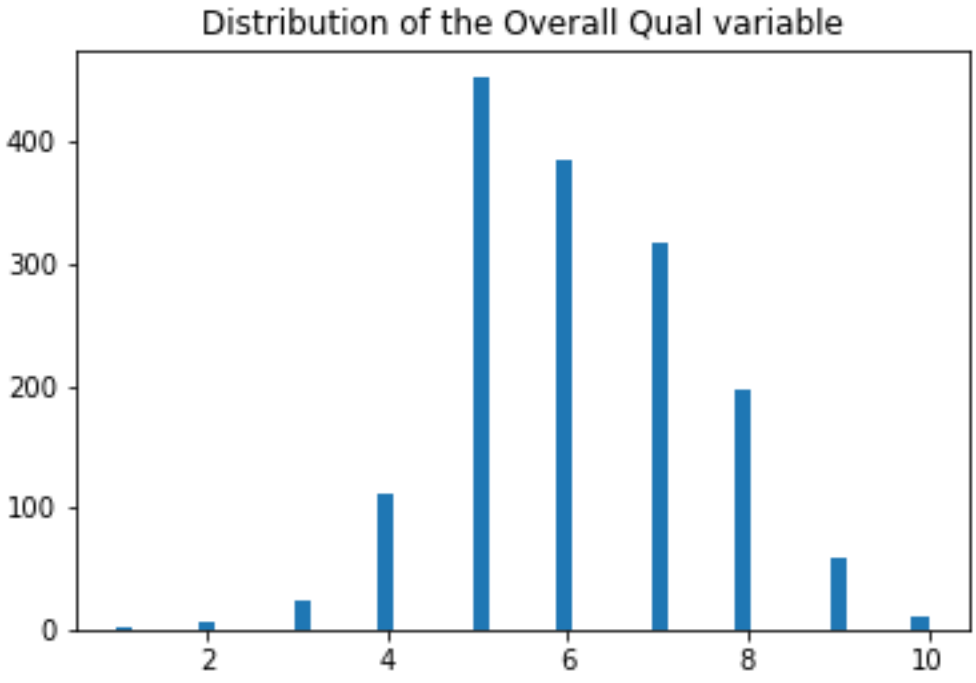

Distribution of the Roof Style variable
1200
1000
800
600
400
200
0
Gable
Hip
Flat
Gambrel
Mansard

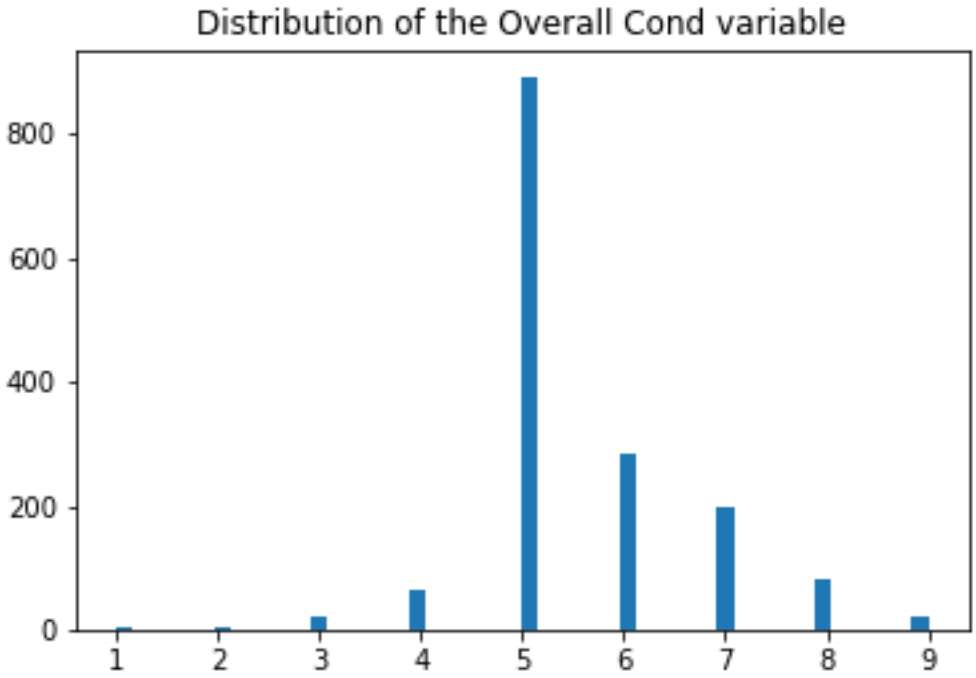

Distribution of the Roof Matl variable
1600
1400
1200
1000
800
600
400
200
0
CompShg
Tar&Grv
Metal
WdShake
WdShngl
Roll

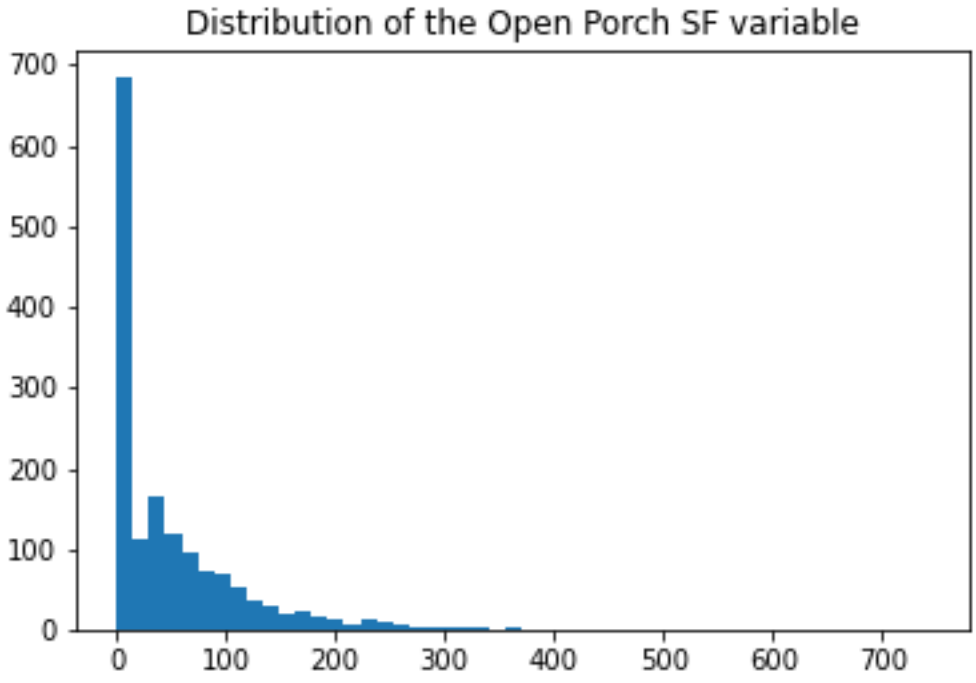

Distribution of the Pool QC variable
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Gd
Fa
Ex
TA

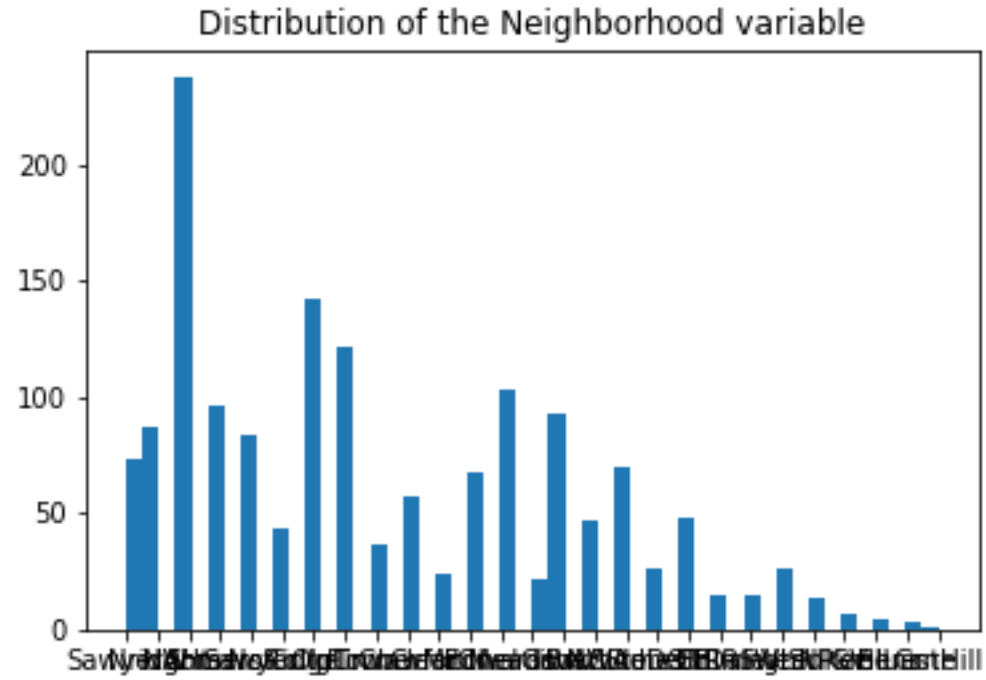

Distribution of the Pool Area variable
1600
1400
1200
1000
800
600
400
200
0
0
100
200
300
400
500
600
700
800

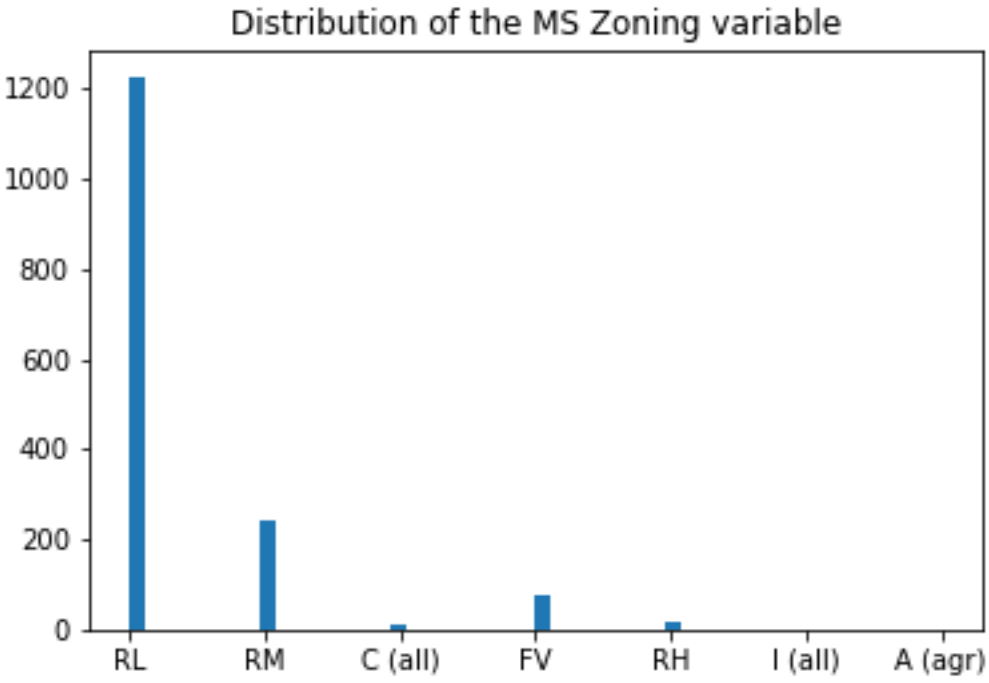

Distribution of the Paved Drive variable
1400
1200
1000
800
600
400
200
0
Y
N
P

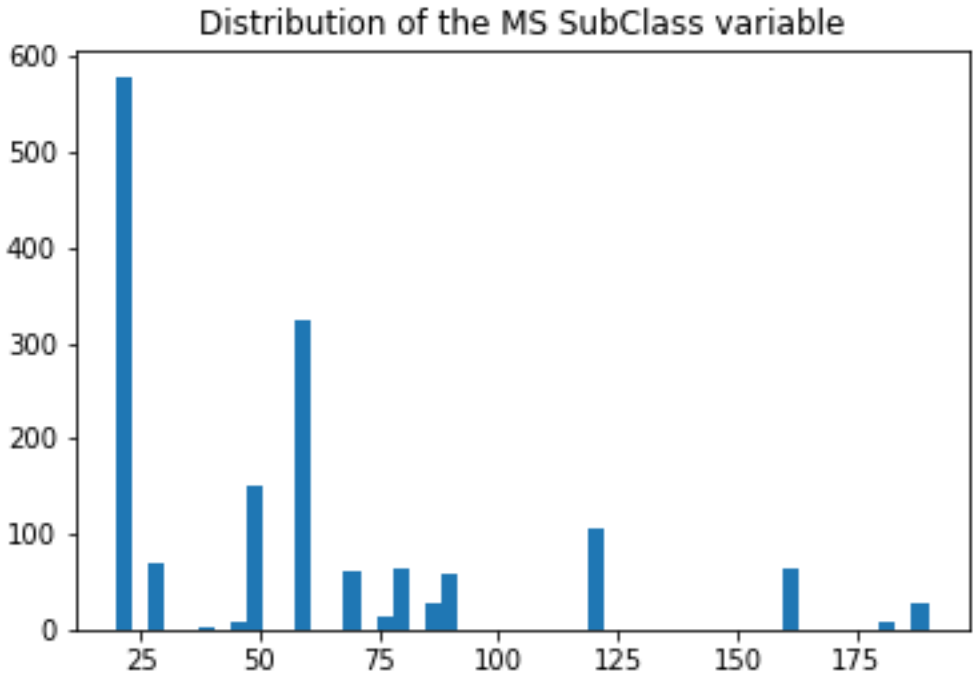

Distribution of the Overall Qual variable
400
300
200
100
0
2
4
6
8
10

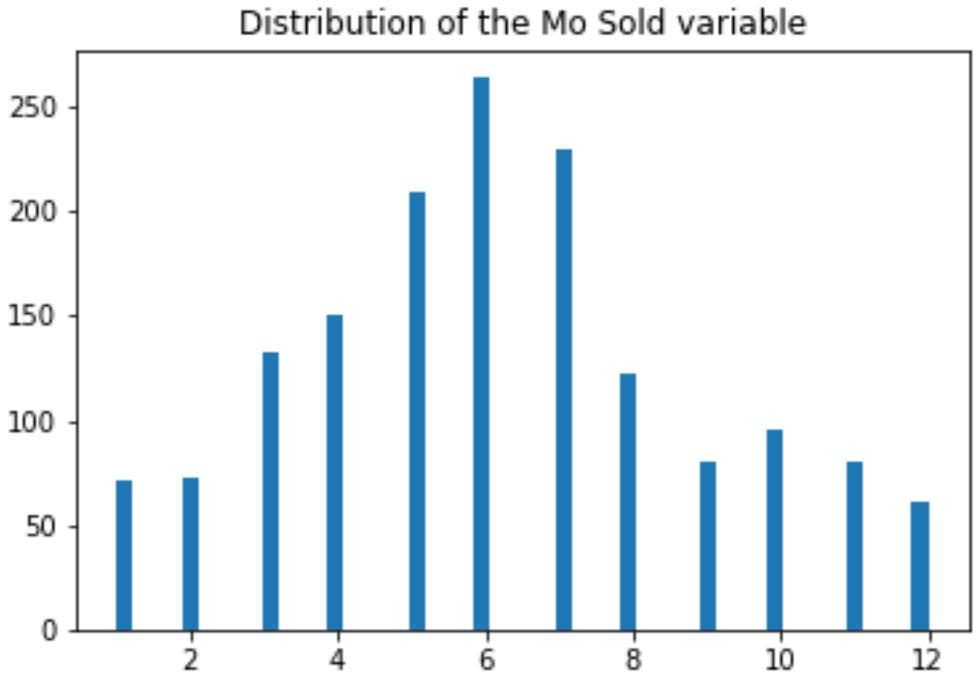

Distribution of the Overall Cond variable
800
600
400
200
0
1
2
3
4
5
6
7
8
9

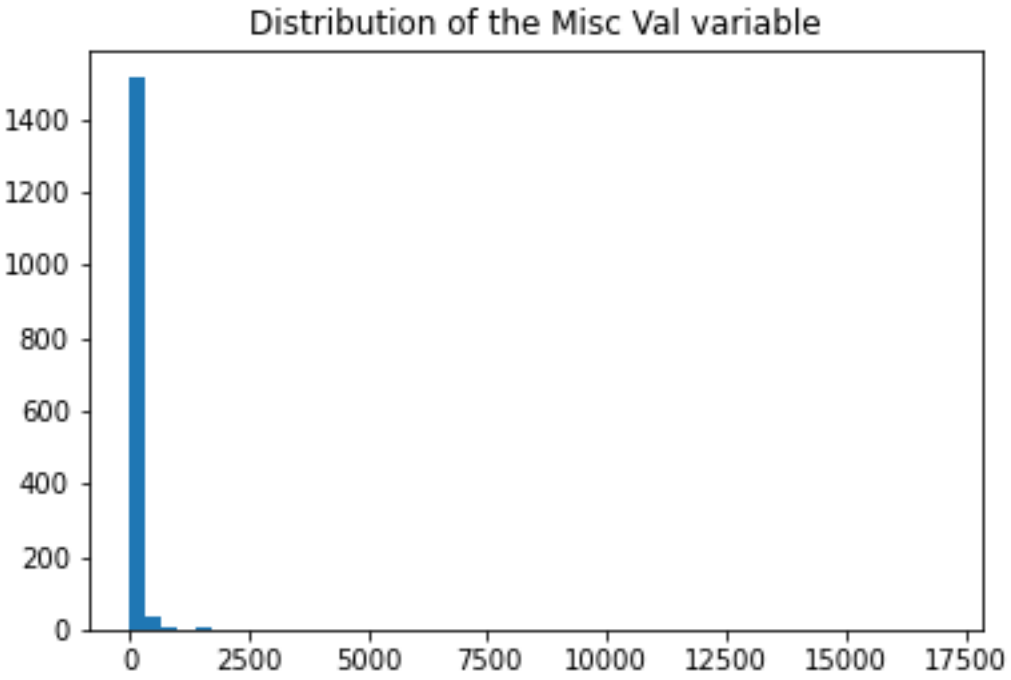

Distribution of the Open Porch SF variable
700
600
500
400
300
200
100
0
0
100
200
300
400
500
600
700

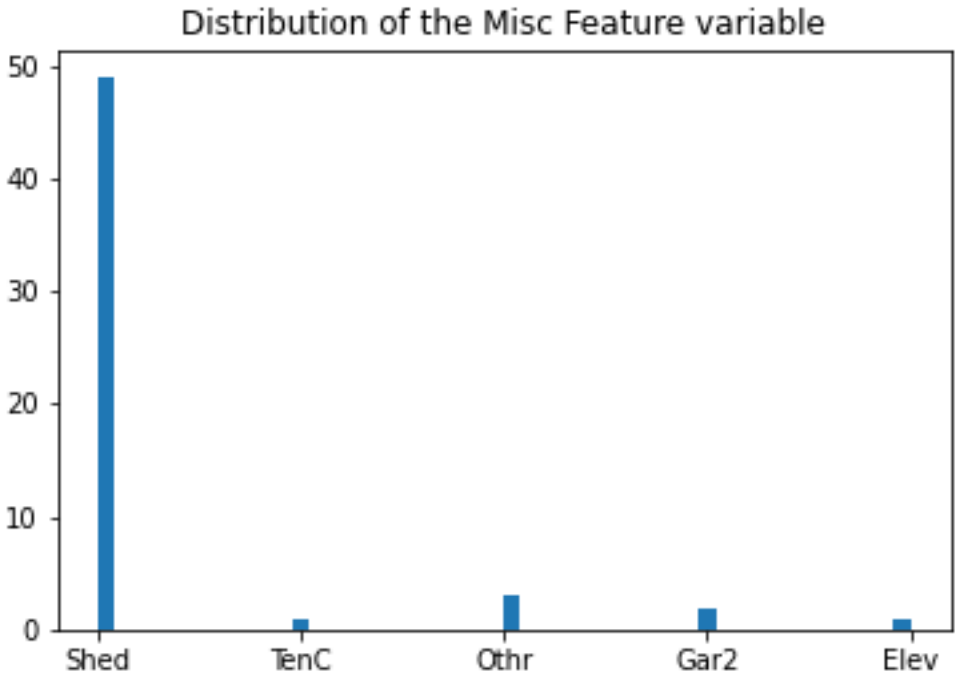

Distribution of the Neighborhood variable
200
150
100
50
0

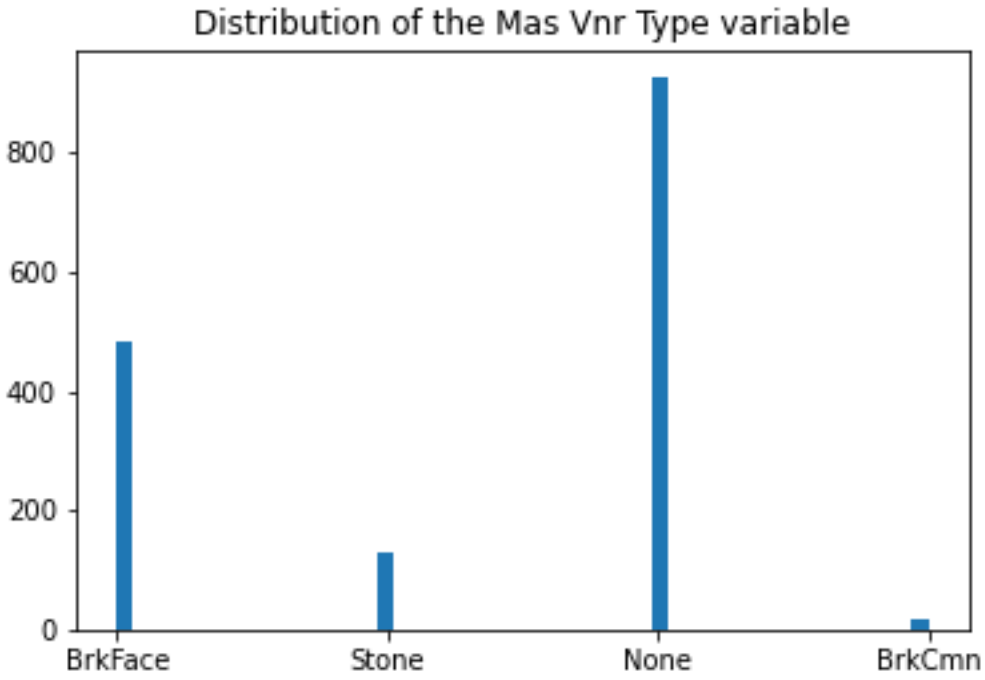

Distribution of the MS Zoning variable
1200
1000
800
600
400
200
0
RL
RM
C (all)
FV
RH
I (all)
A (agr)

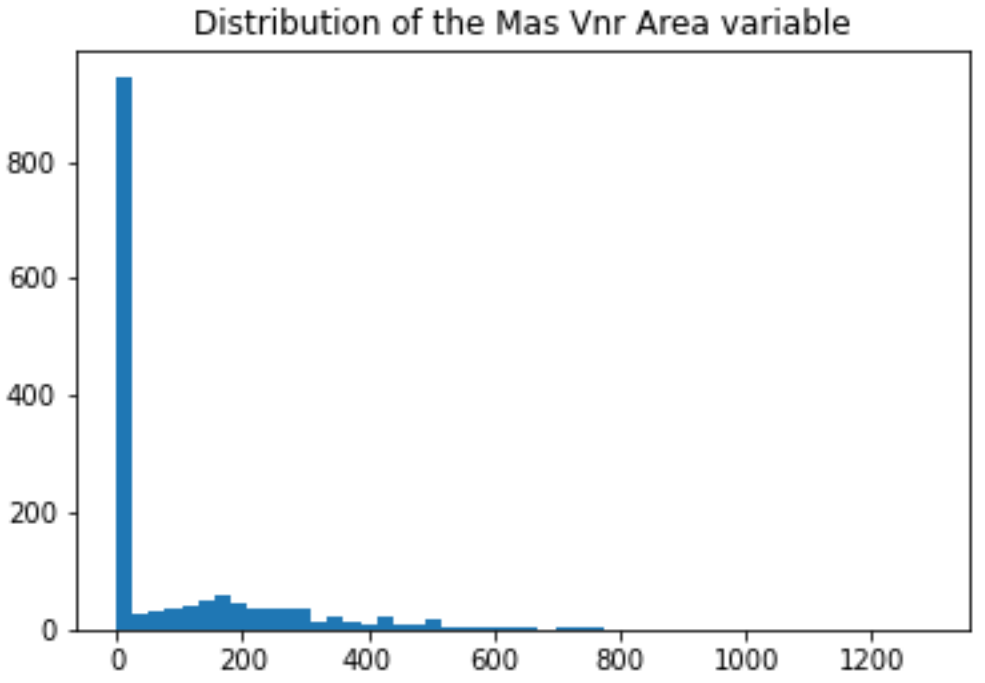

Distribution of the MS SubClass variable
600
500
400
300
200
100
0
25
50
75
100
125
150
175

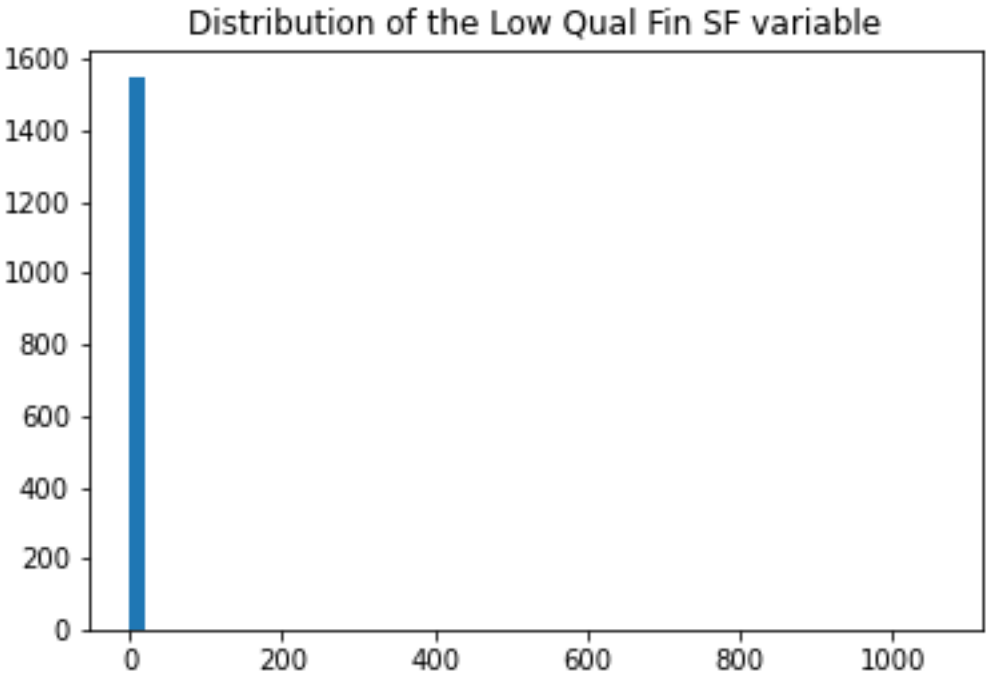

Distribution of the Mo Sold variable
250
200
150
100
50
0
2
4
6
8
10
12

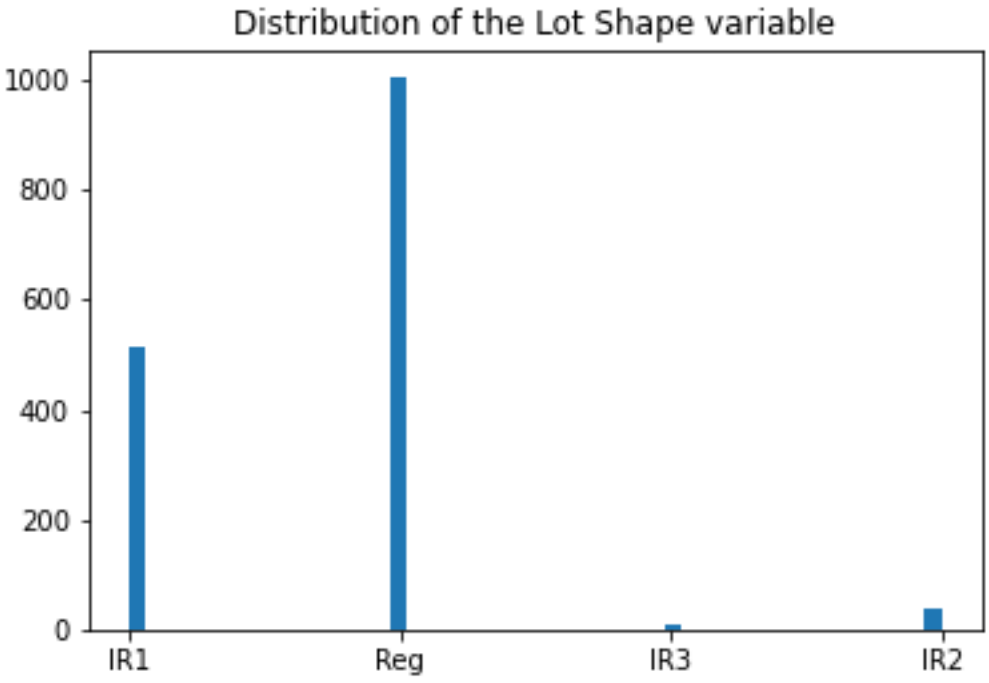

Distribution of the Misc Val variable
1400
1200
1000
800
600
400
200
0
0
2500
5000
7500
10000
12500
15000
17500

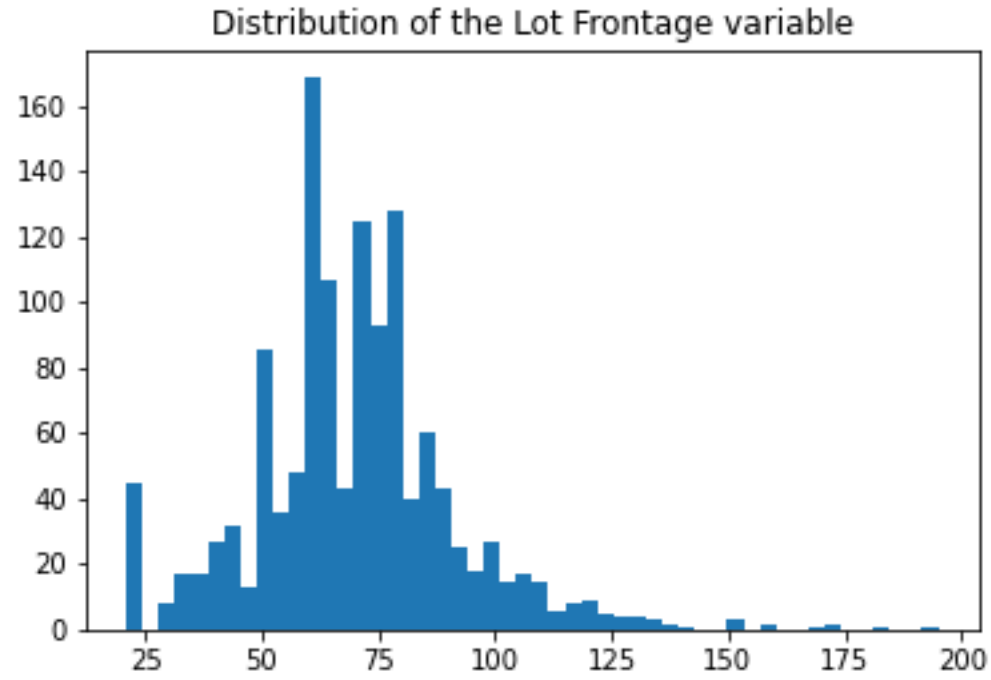

Distribution of the Misc Feature variable
50
40
30
20
10
0
Shed
TenC
Othr
Gar2
Elev

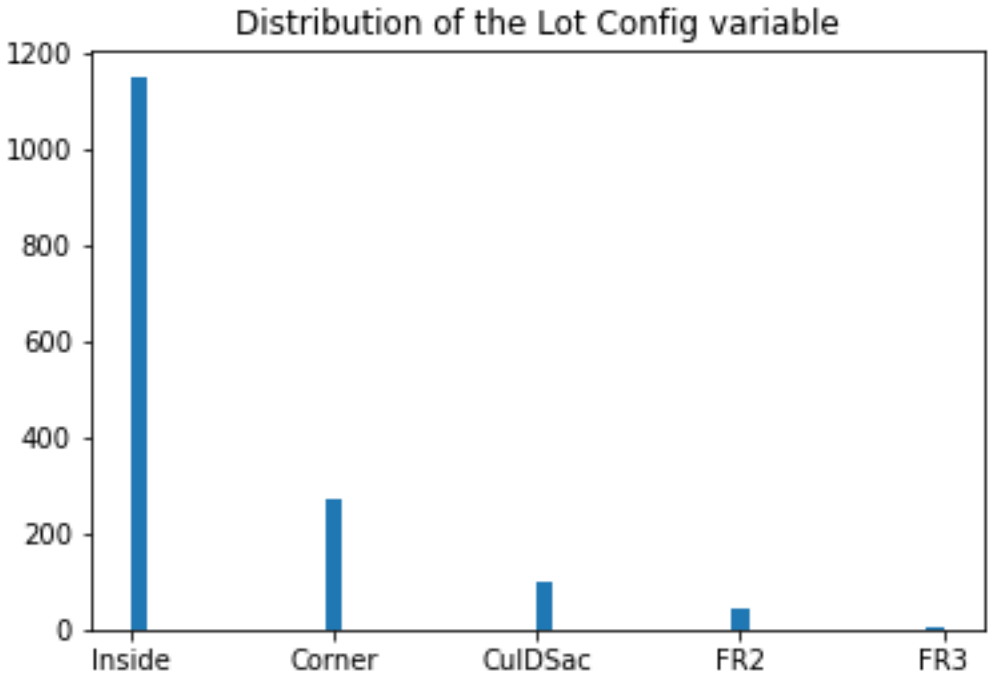

Distribution of the Mas Vnr Type variable
800
600
400
200
0
BrkFace
Stone
None
BrkCmn

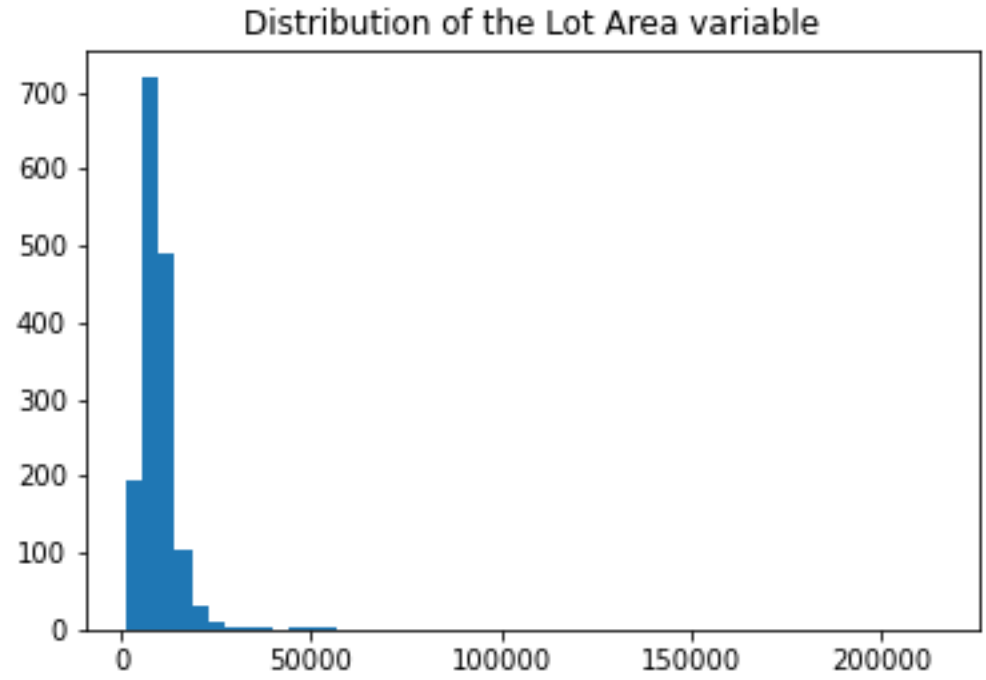

Distribution of the Mas Vnr Area variable
800
600
400
200
0
0
200
400
600
800
1000
1200

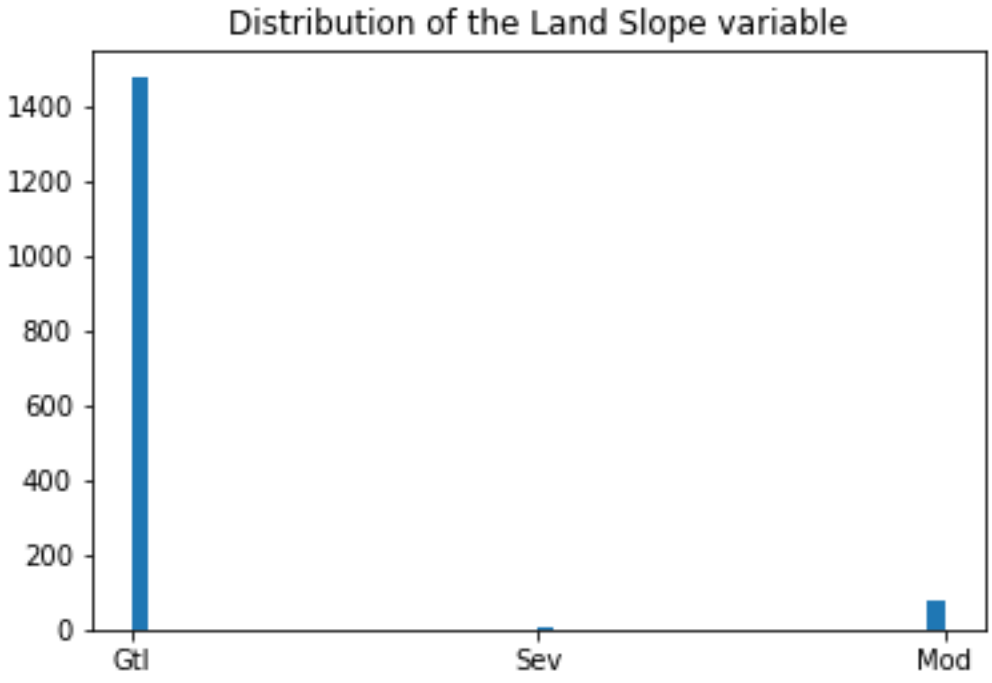

Distribution of the Low Qual Fin SF variable
1600
1400
1200
1000
800
600
400
200
0
0
200
400
600
800
1000

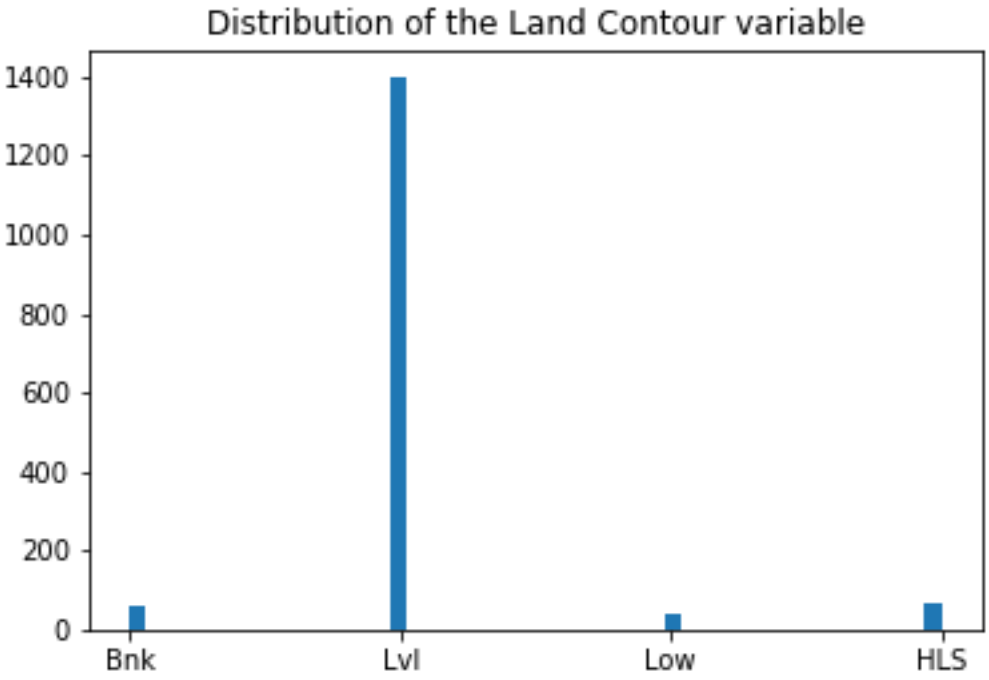

Distribution of the Lot Shape variable
1000
800
600
400
200
0
IR1
Reg
IR3
IR2

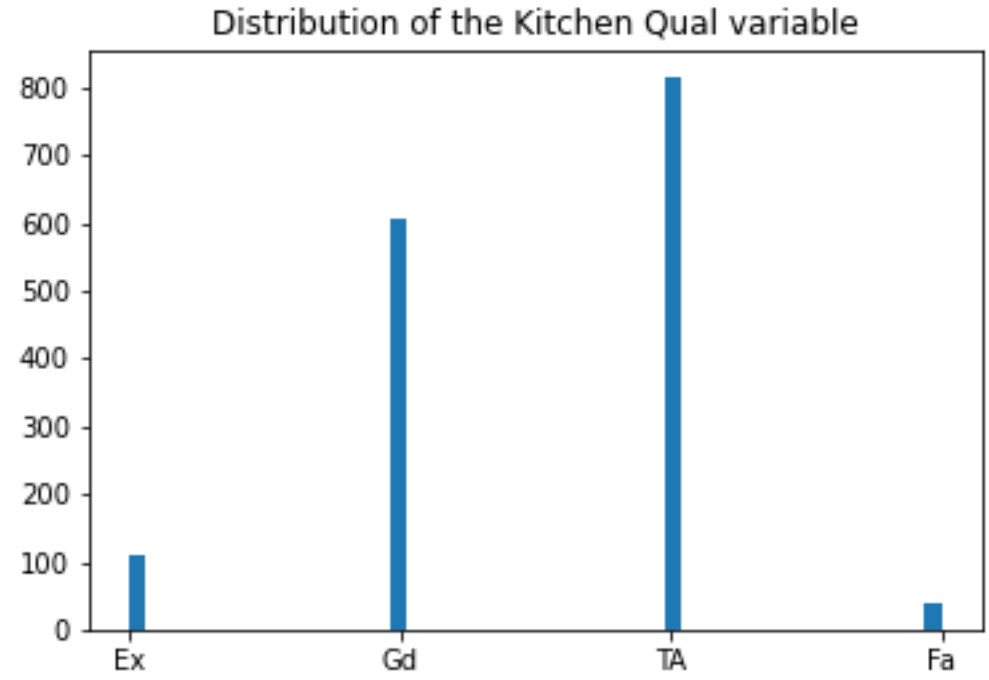

Distribution of the Lot Frontage variable
160
140
120
100
80
60
40
20
0
25
50
75
100
125
150
175
200

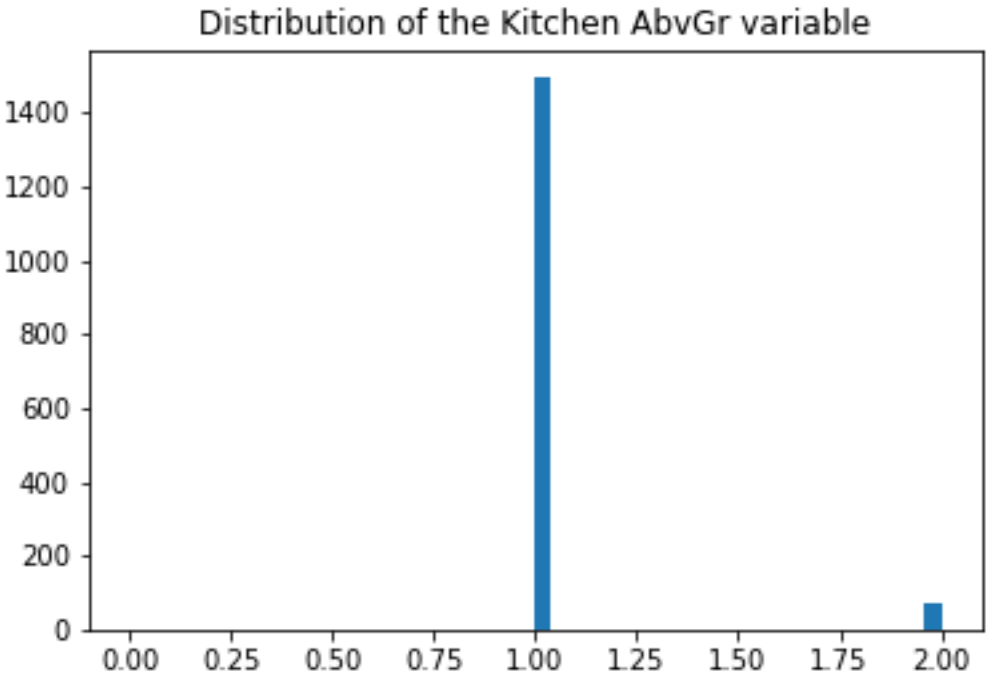

Distribution of the Lot Config variable
1200
1000
800
600
400
200
0
Inside
Corner
CulDSac
FR2
FR3

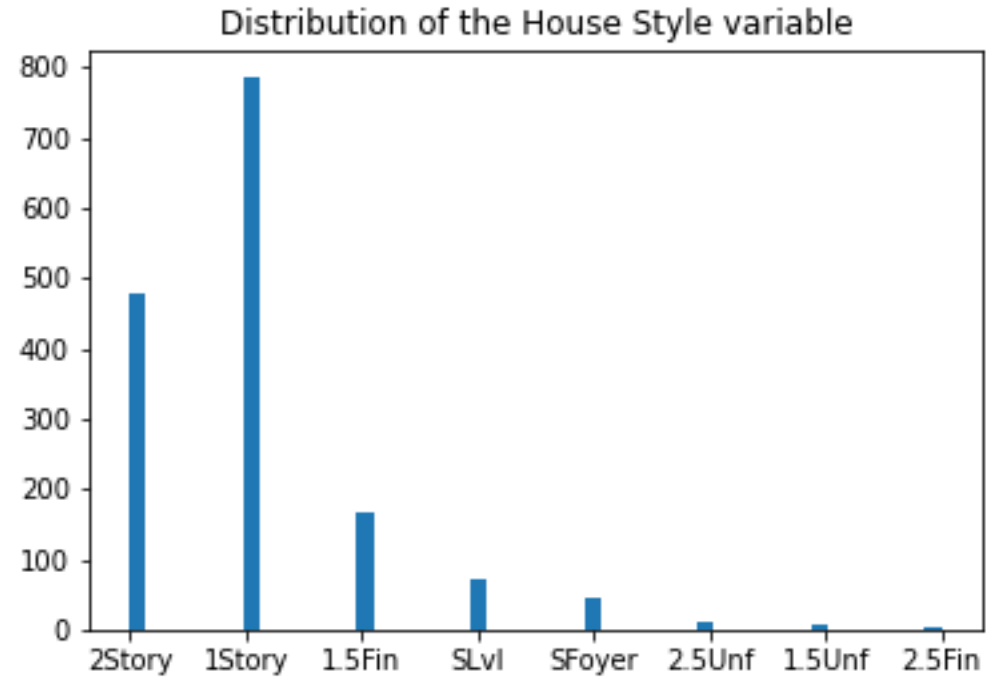

Distribution of the Lot Area variable
700
600
500
400
300
200
100
0
0
50000
100000
150000
200000

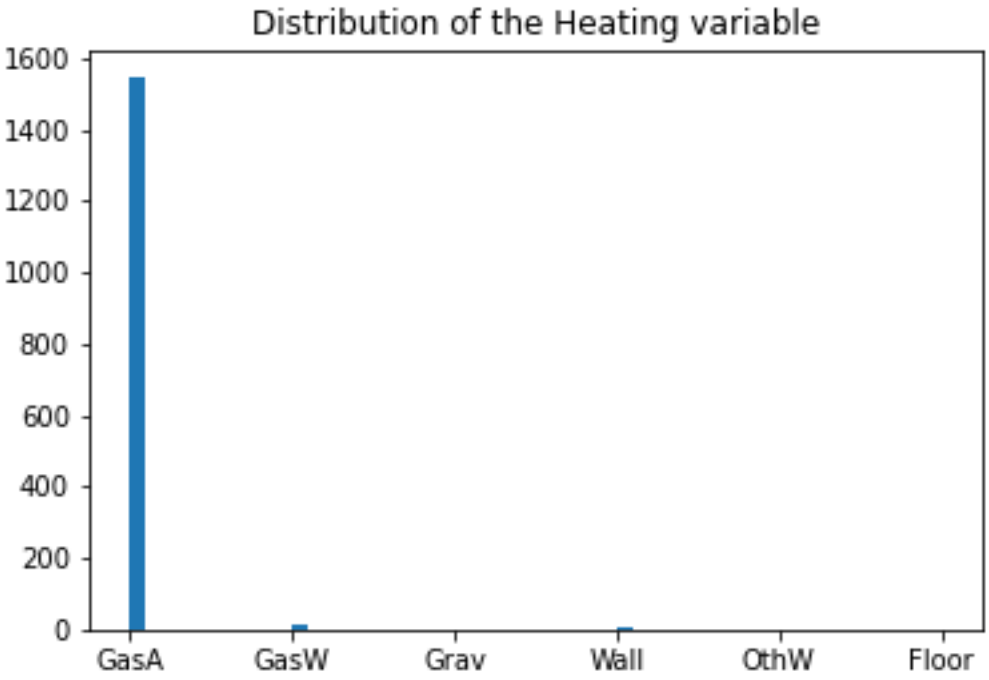

Distribution of the Land Slope variable
1400
1200
1000
800
600
400
200
0
Gtl
Sev
Mod

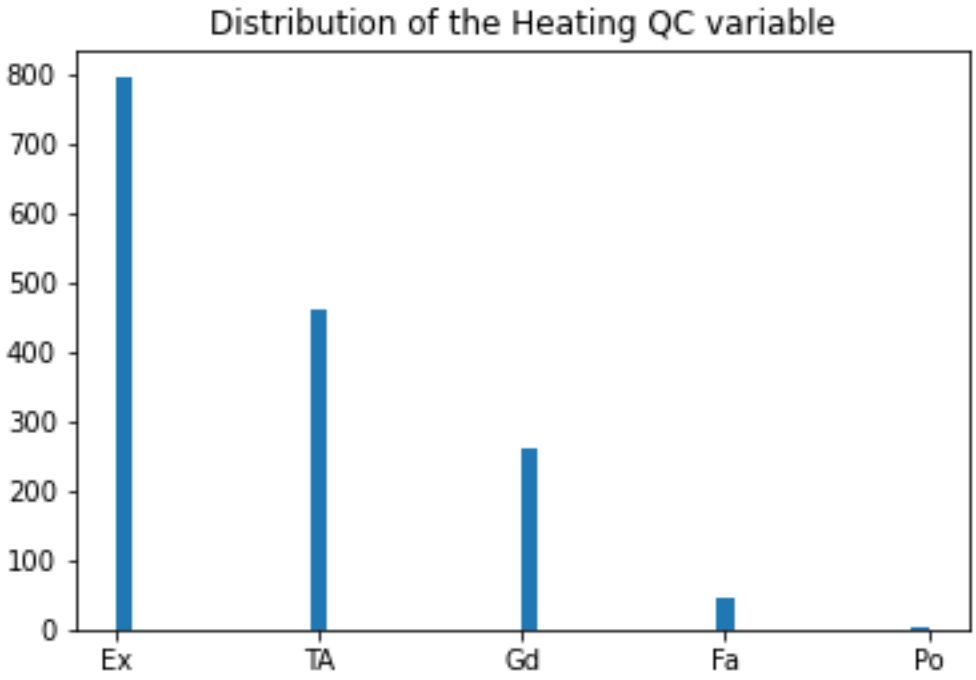

Distribution of the Land Contour variable
1400
1200
1000
800
600
400
200
0
Bnk
Lvl
Low
HLS

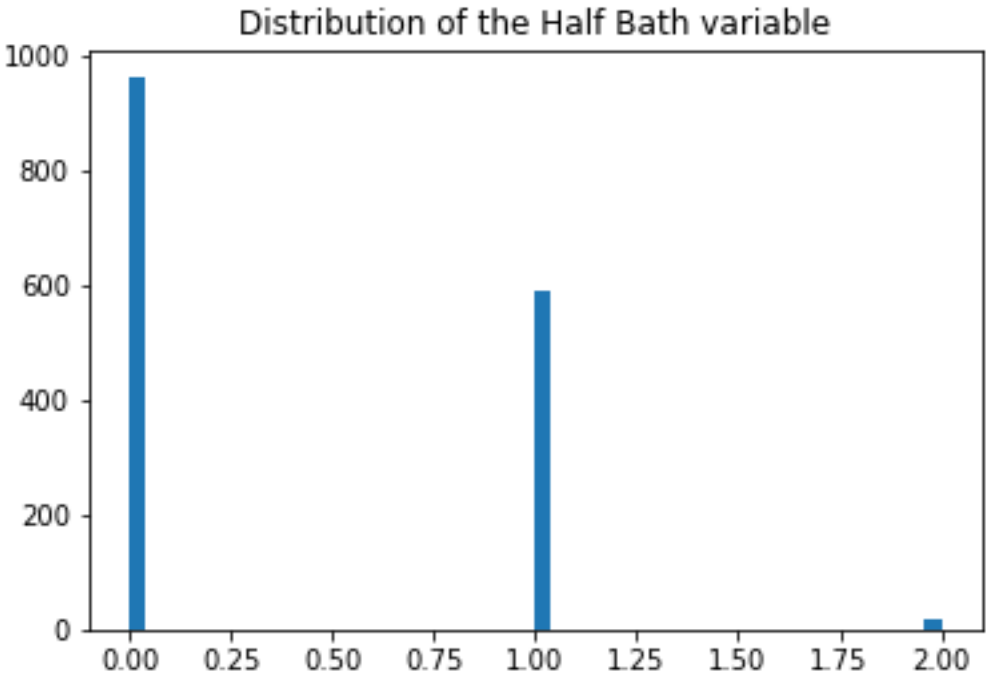

Distribution of the Kitchen Qual variable
800
700
600
500
400
300
200
100
0
Ex
Gd
TA
Fa

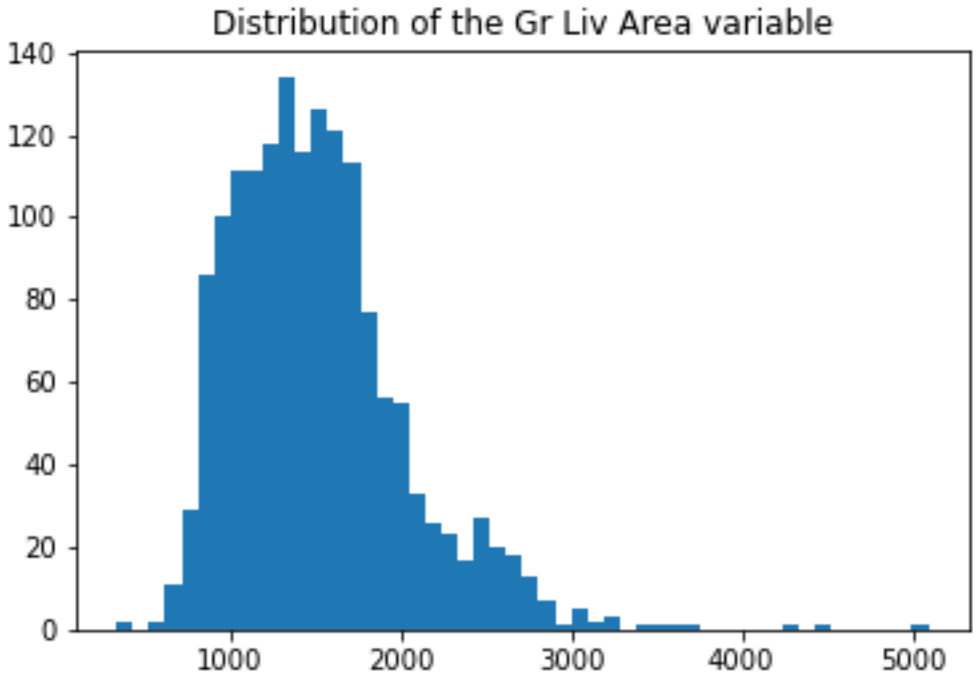

Distribution of the Kitchen AbvGr variable
1400
1200
1000
800
600
400
200
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

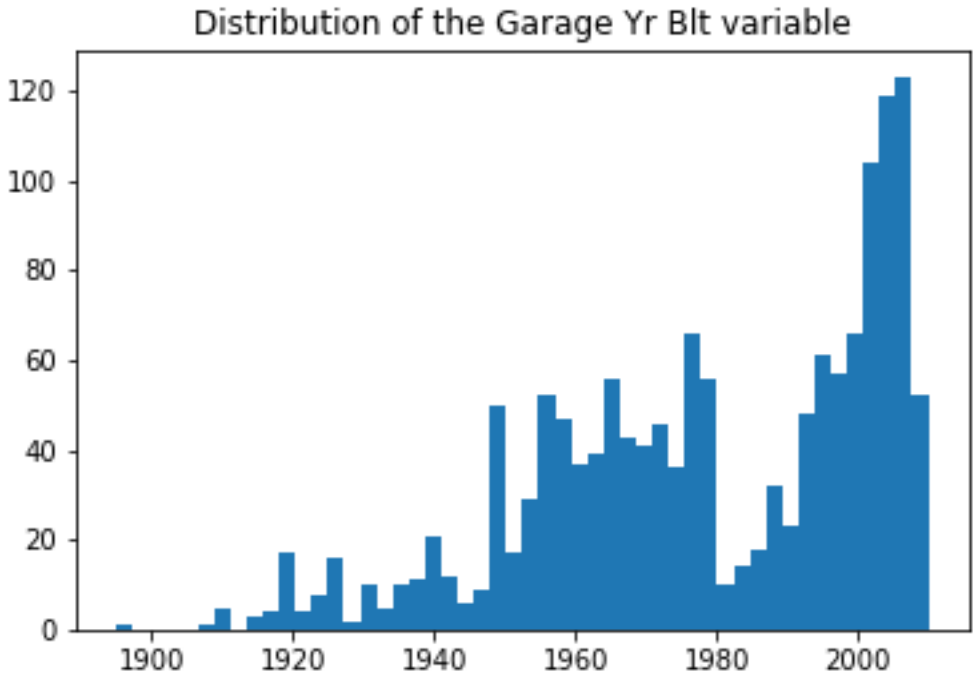

Distribution of the House Style variable
800
700
600
500
400
300
200
100
0
2Story
1Story
1.5Fin
SLvl
SFoyer
2.5Unf
1.5Unf
2.5Fin

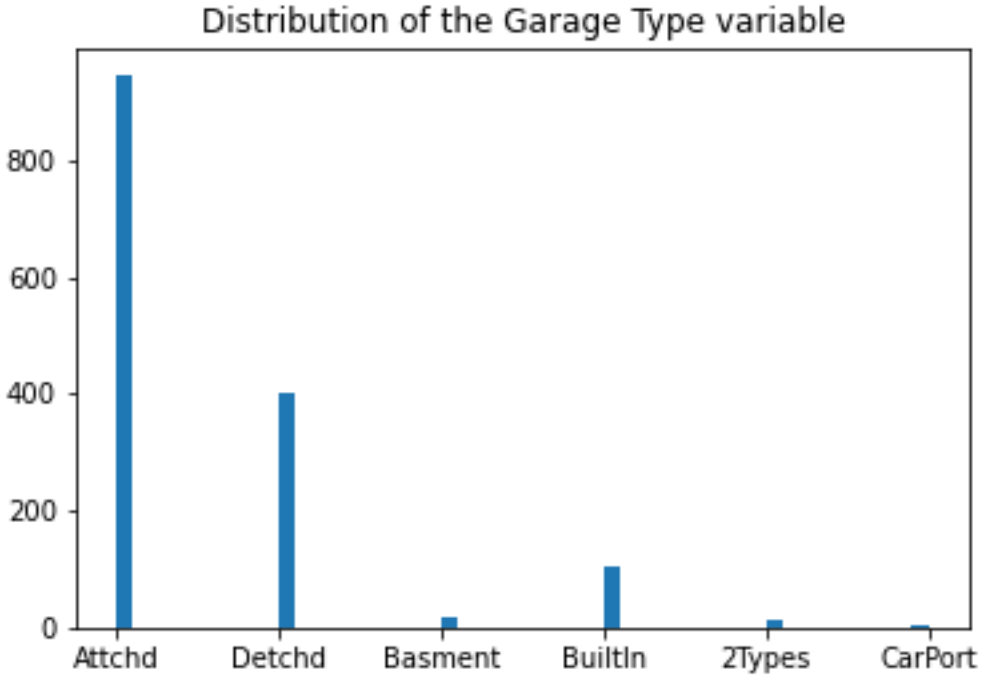

Distribution of the Heating variable
1600
1400
1200
1000
800
600
400
200
0
GasA
GasW
Grav
Wall
OthW
Floor

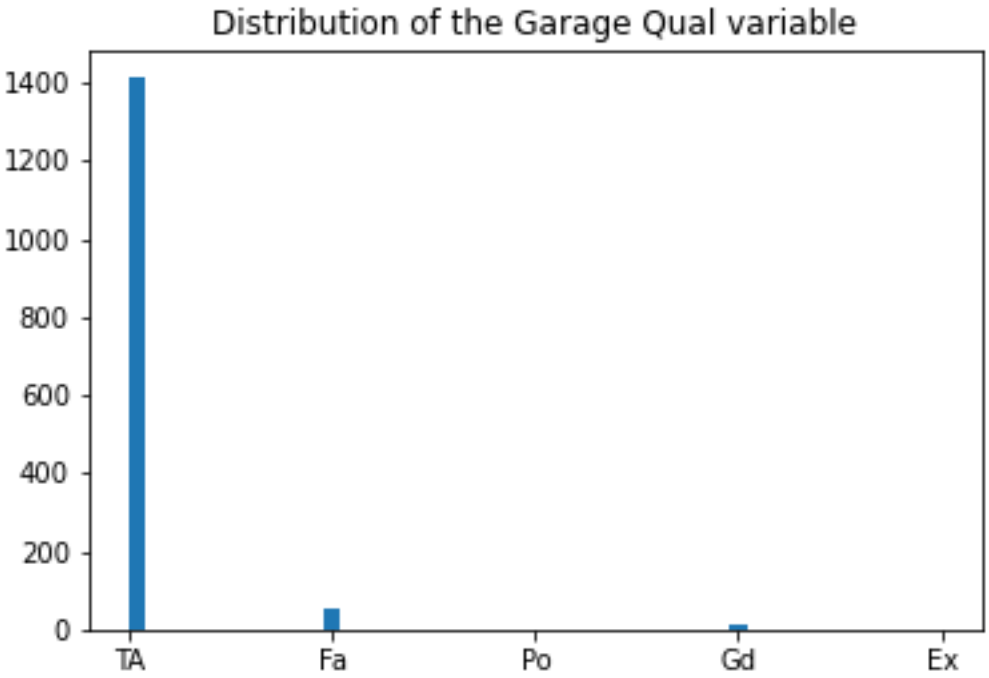

Distribution of the Heating QC variable
800
700
600
500
400
300
200
100
0
Ex
TA
Gd
Fa
Po

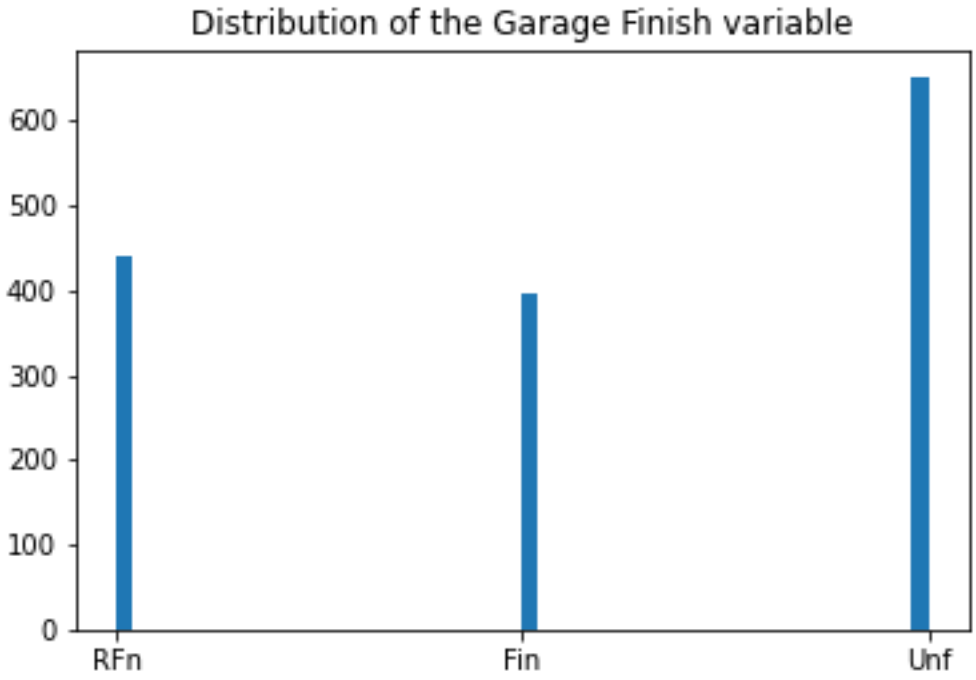

Distribution of the Half Bath variable
1000
800
600
400
200
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

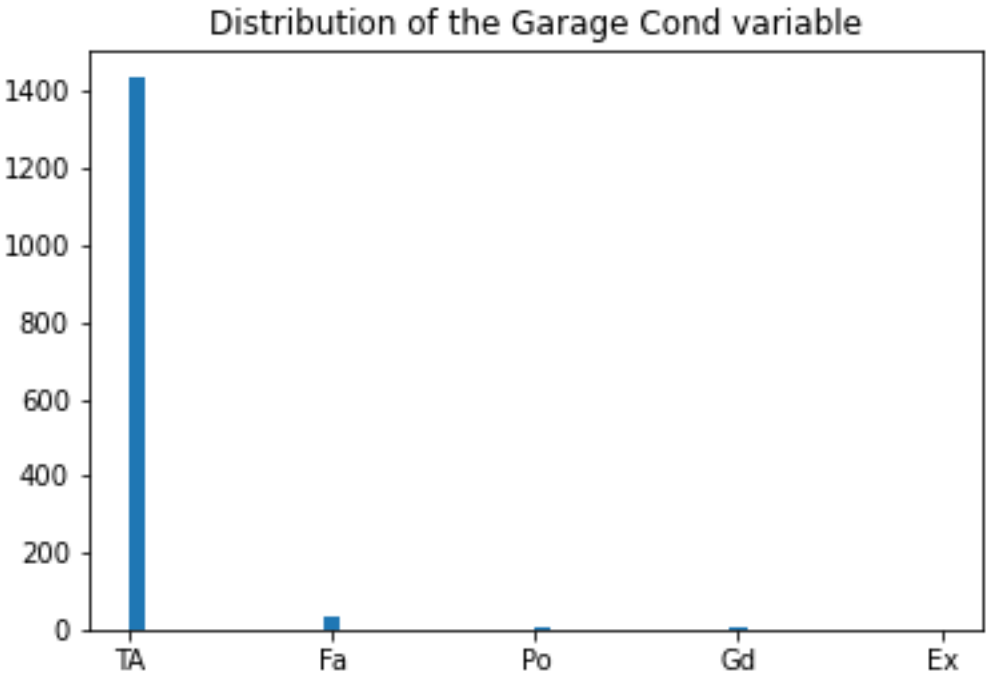

Distribution of the Gr Liv Area variable
140
120
100
80
60
40
20
0
1000
2000
3000
4000
5000

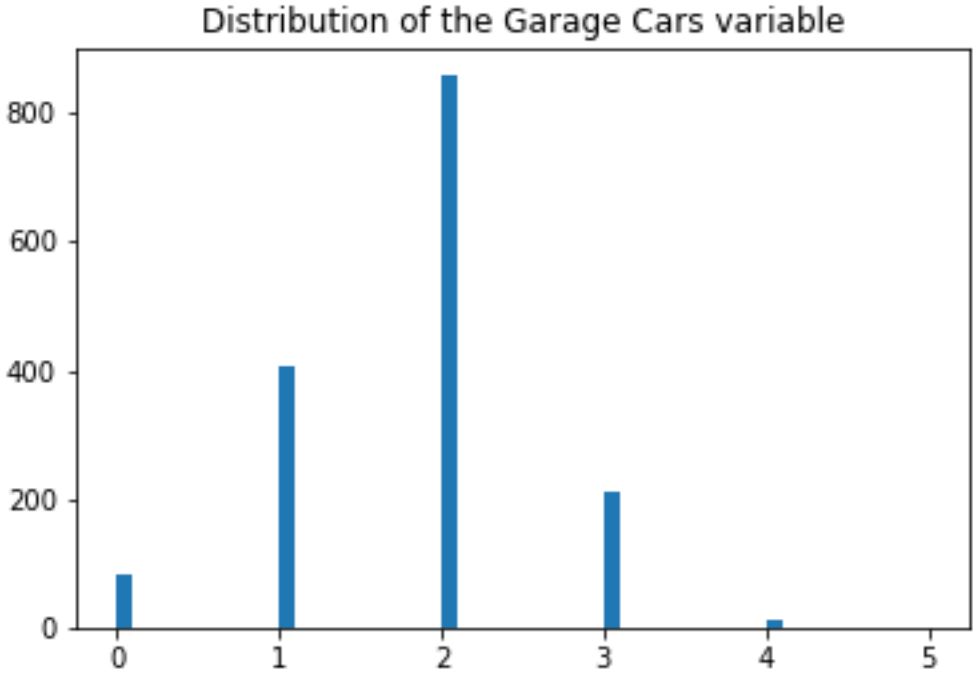

Distribution of the Garage Yr Blt variable
120
100
80
60
40
20
0
1900
1920
1940
1960
1980
2000

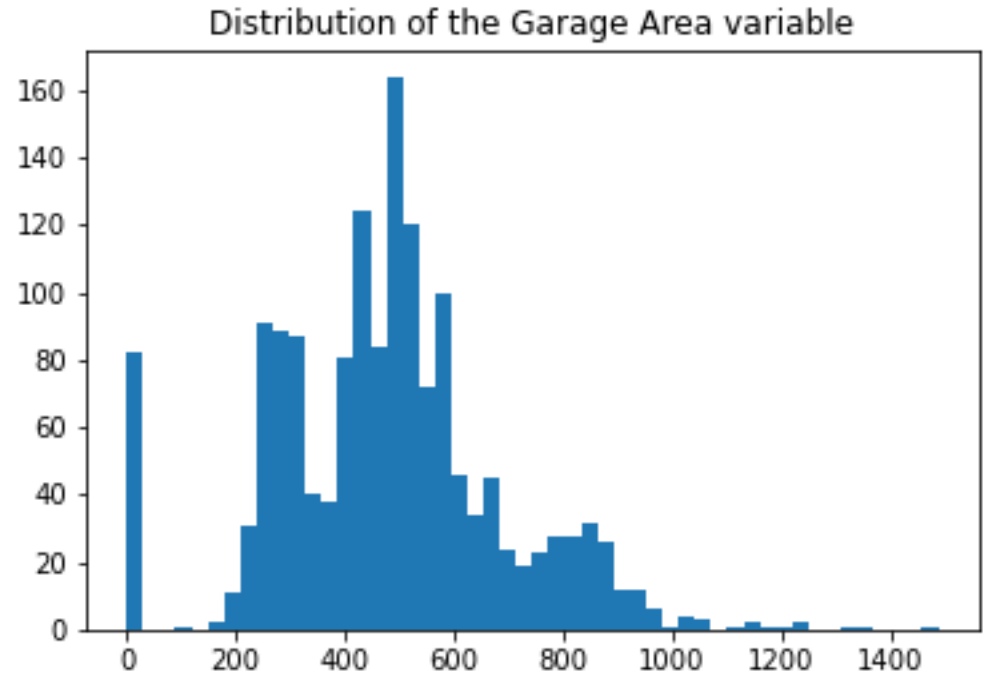

Distribution of the Garage Type variable
800
600
400
200
0
Attchd
Detchd
Basment
BuiltIn
2Types
CarPort

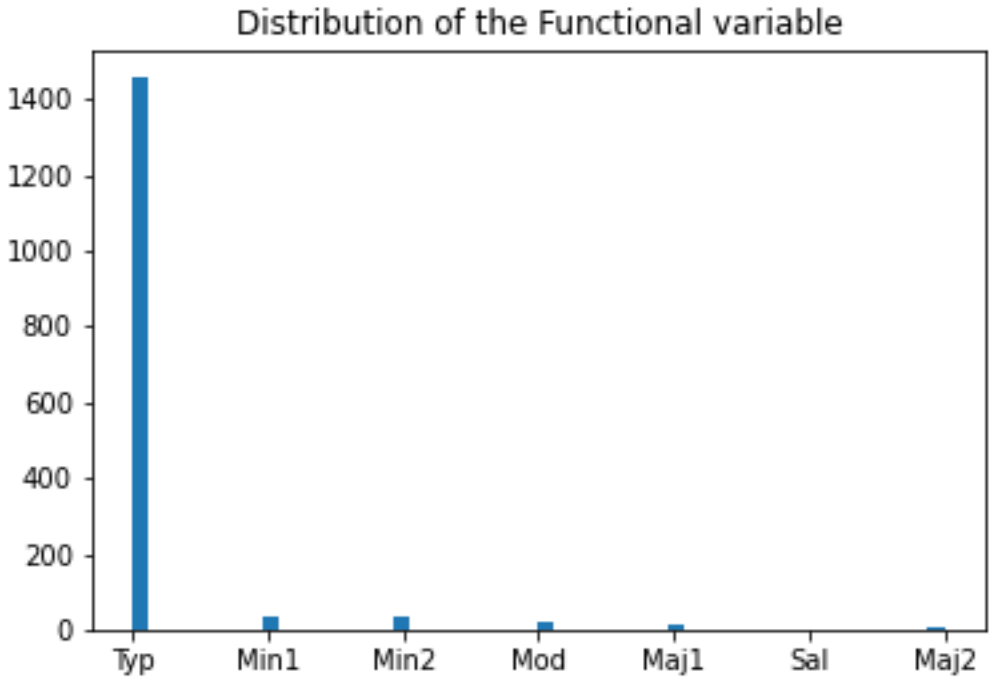

Distribution of the Garage Qual variable
1400
1200
1000
800
600
400
200
0
TA
Fa
Po
Gd
Ex

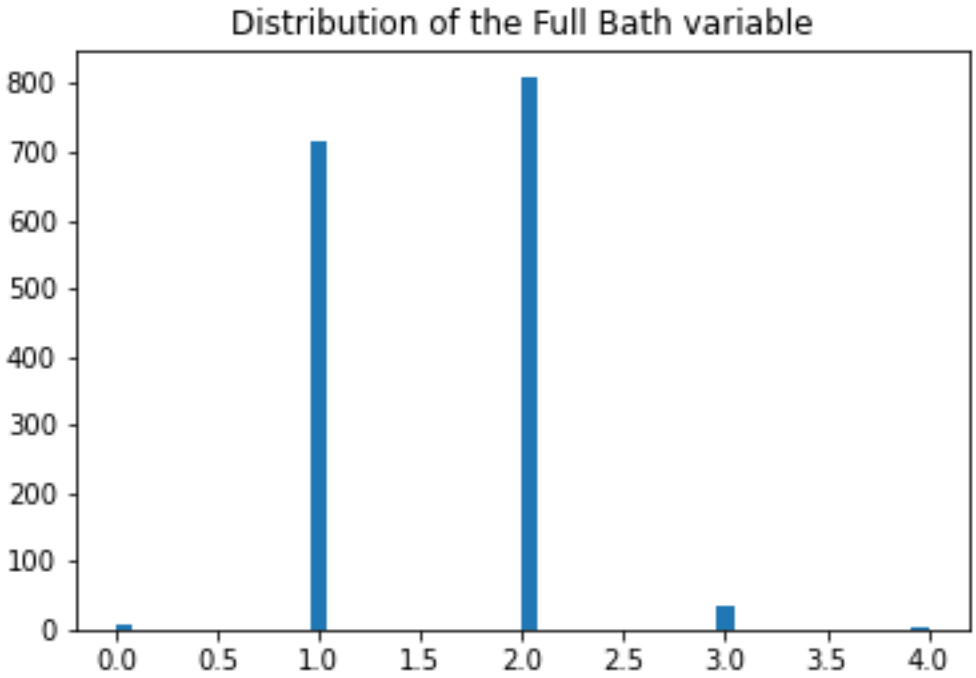

Distribution of the Garage Finish variable
600
500
400
300
200
100
0
RFn
Fin
Unf

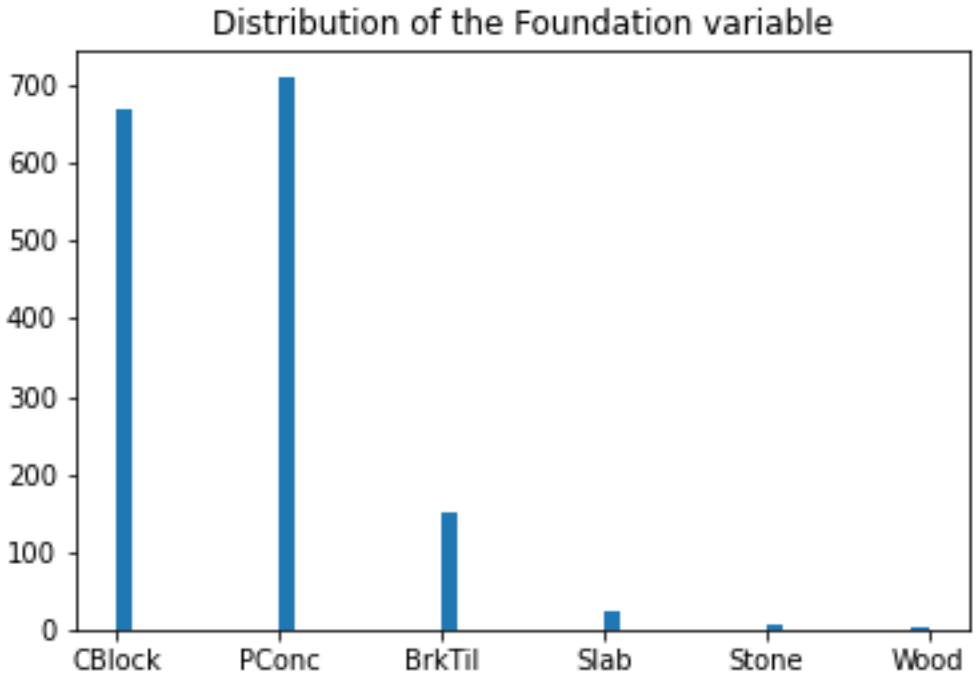

Distribution of the Garage Cond variable
1400
1200
1000
800
600
400
200
0
TA
Fa
Po
Gd
Ex

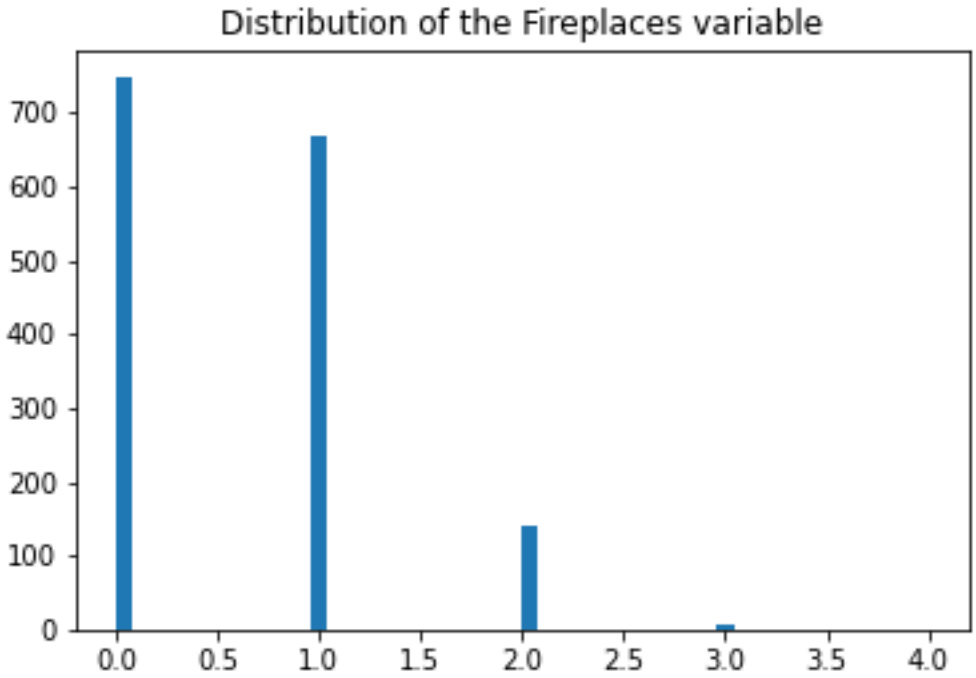

Distribution of the Garage Cars variable
800
600
400
200
0
0
1
2
3
4
5

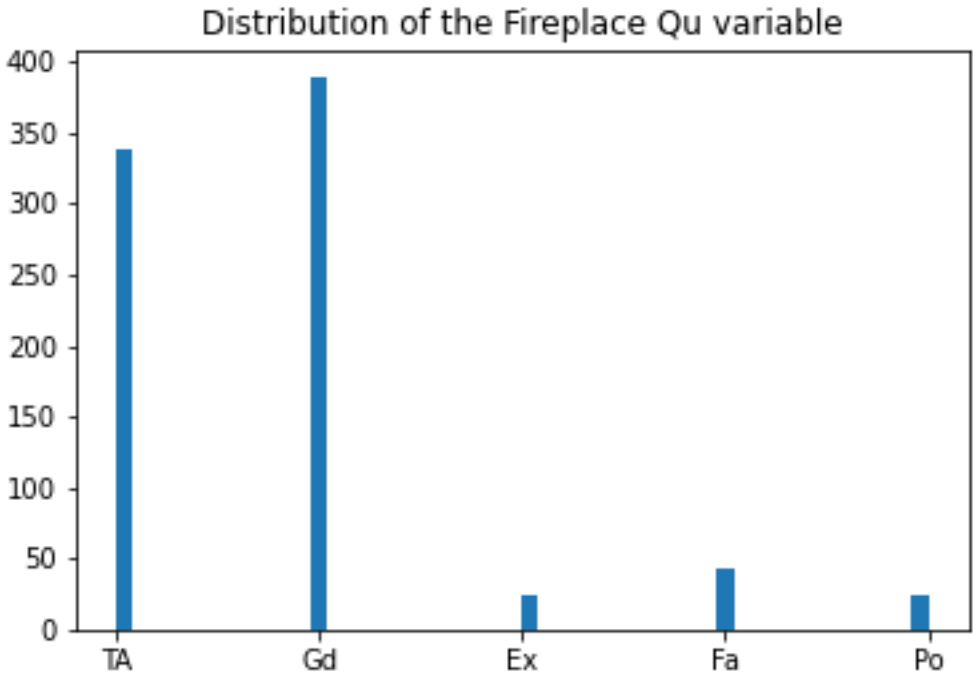

Distribution of the Garage Area variable
160
140
120
100
80
60
40
20
0
0
200
400
600
800
1000
1200
1400

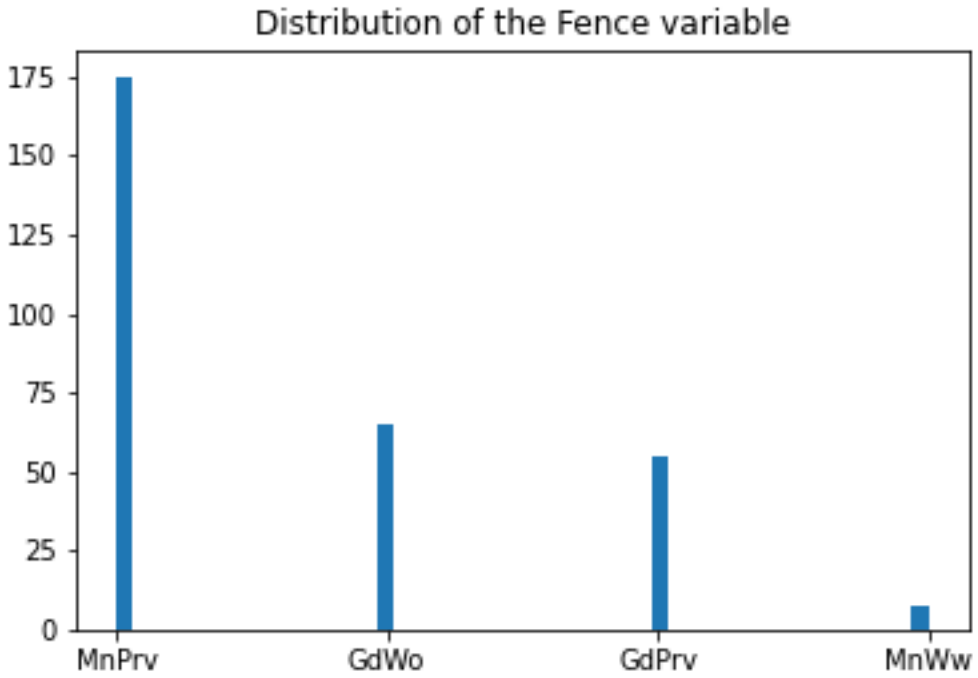

Distribution of the Functional variable
1400
1200
1000
800
600
400
200
0
Typ
Min1
Min2
Mod
Maj1
Sal
Maj2

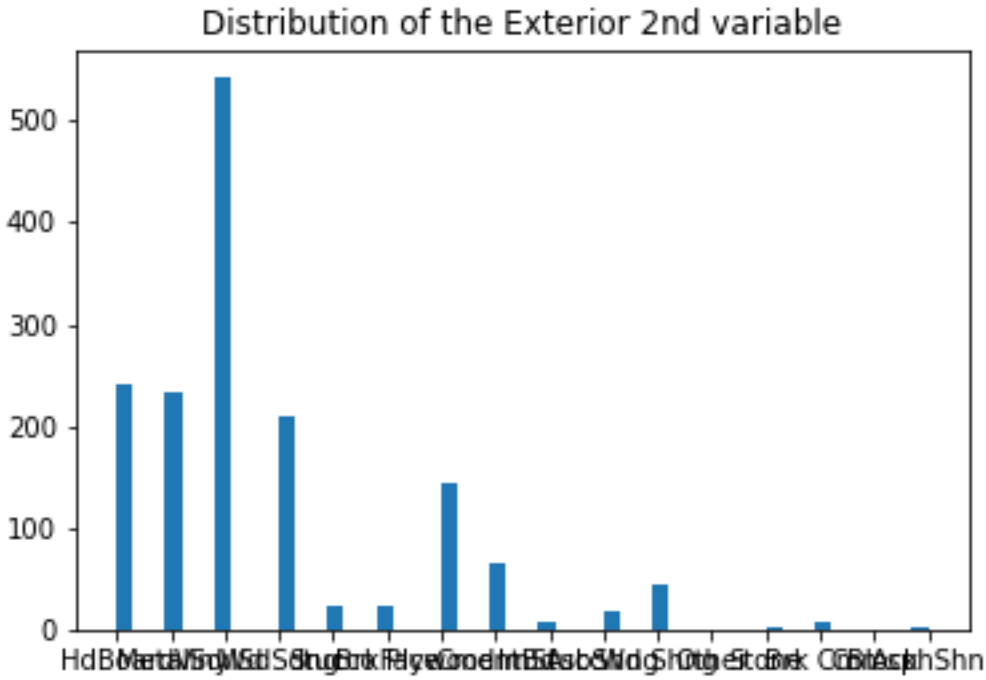

Distribution of the Full Bath variable
800
700
600
500
400
300
200
100
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

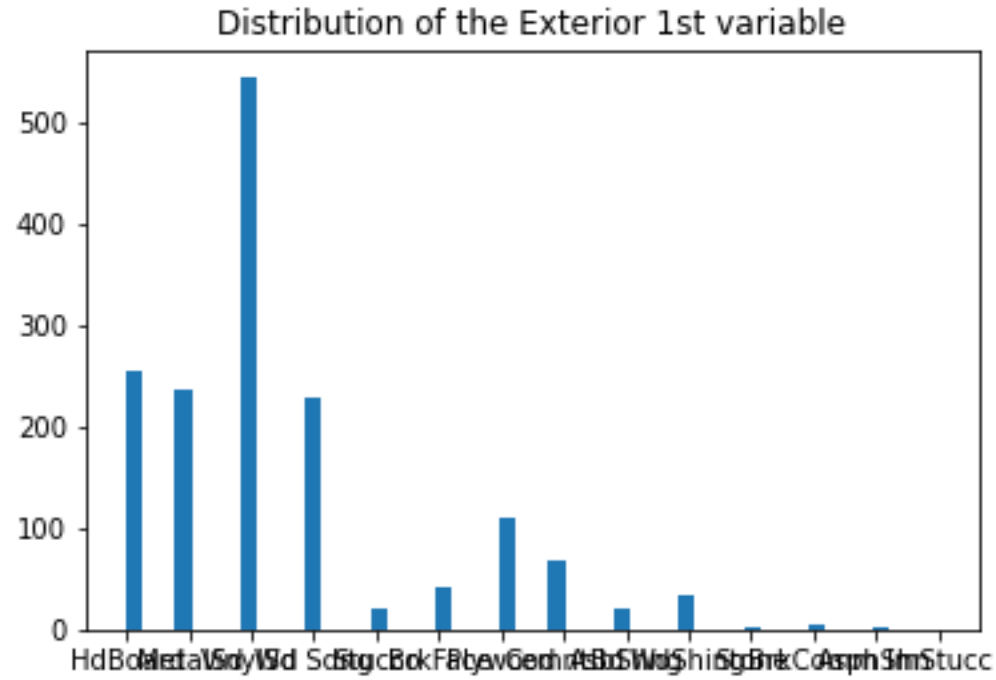

Distribution of the Foundation variable
700
600
500
400
300
200
100
0
CBlock
PConc
BrkTil
Slab
Stone
Wood

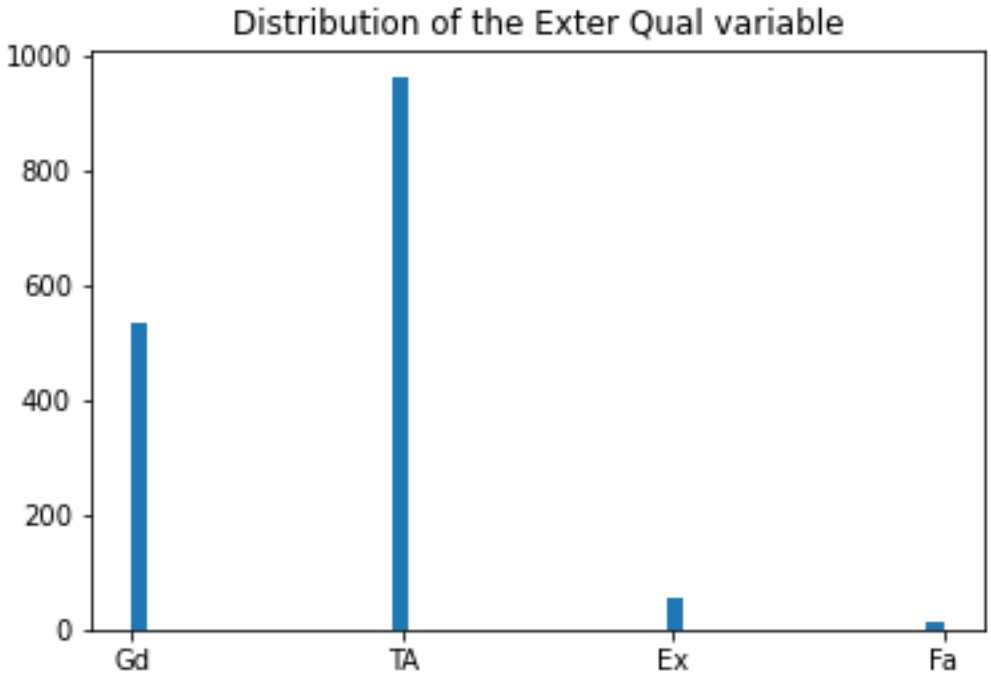

Distribution of the Fireplaces variable
700
600
500
400
300
200
100
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

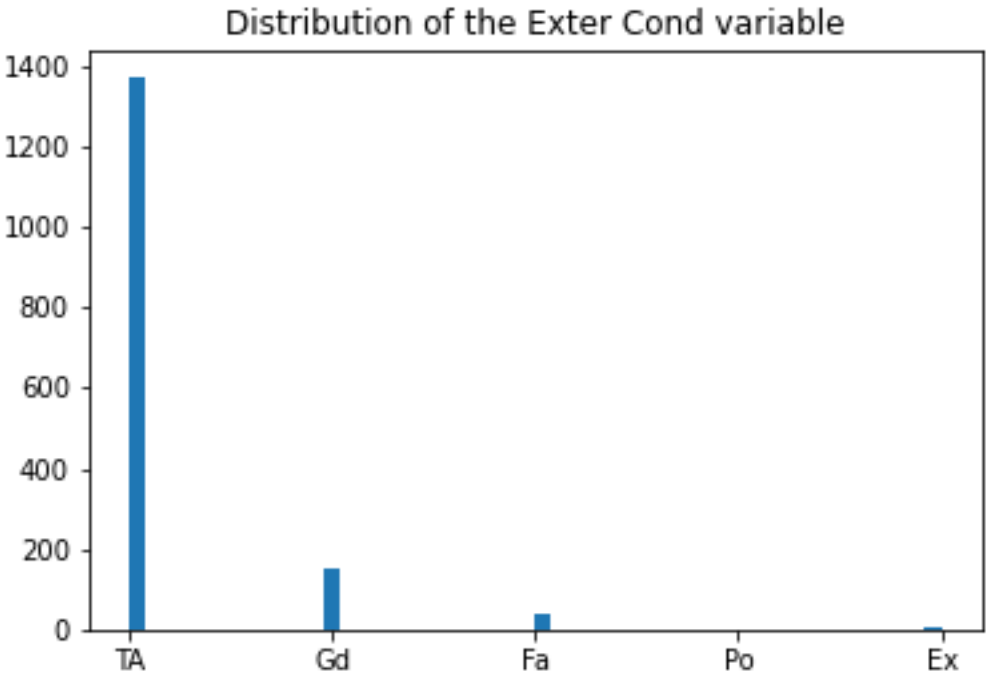

Distribution of the Fireplace Qu variable
400
350
300
250
200
150
100
50
0
TA
Gd
Ex
Fa
Po

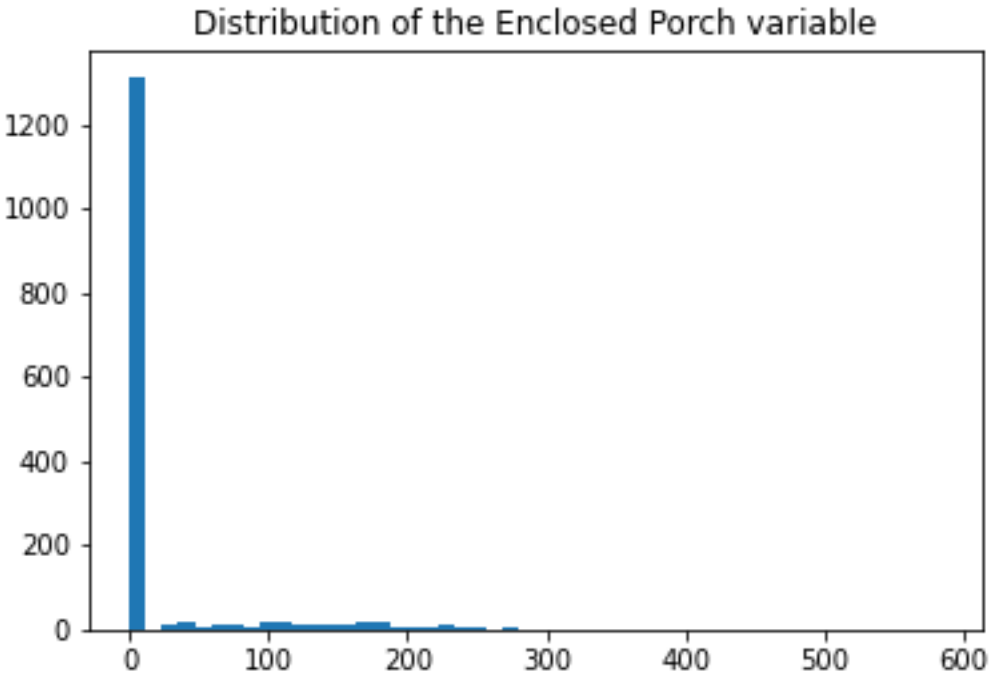

Distribution of the Fence variable
175
150
125
100
75
50
25
0
MnPrv
GdWo
GdPrv
MnWw

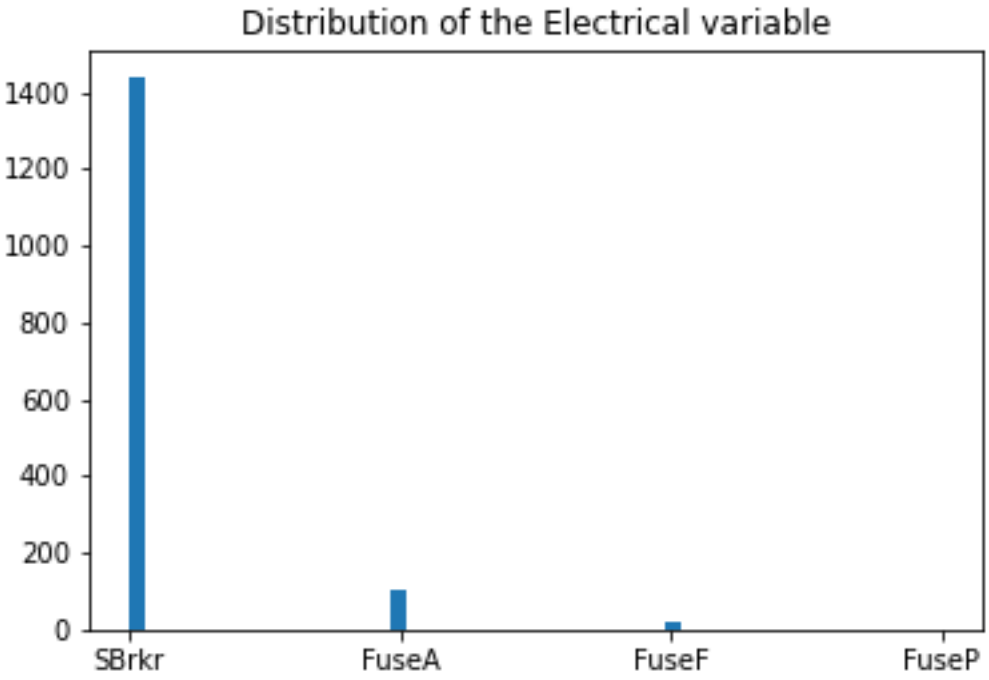

Distribution of the Exterior 2nd variable
500
400
300
200
100
0

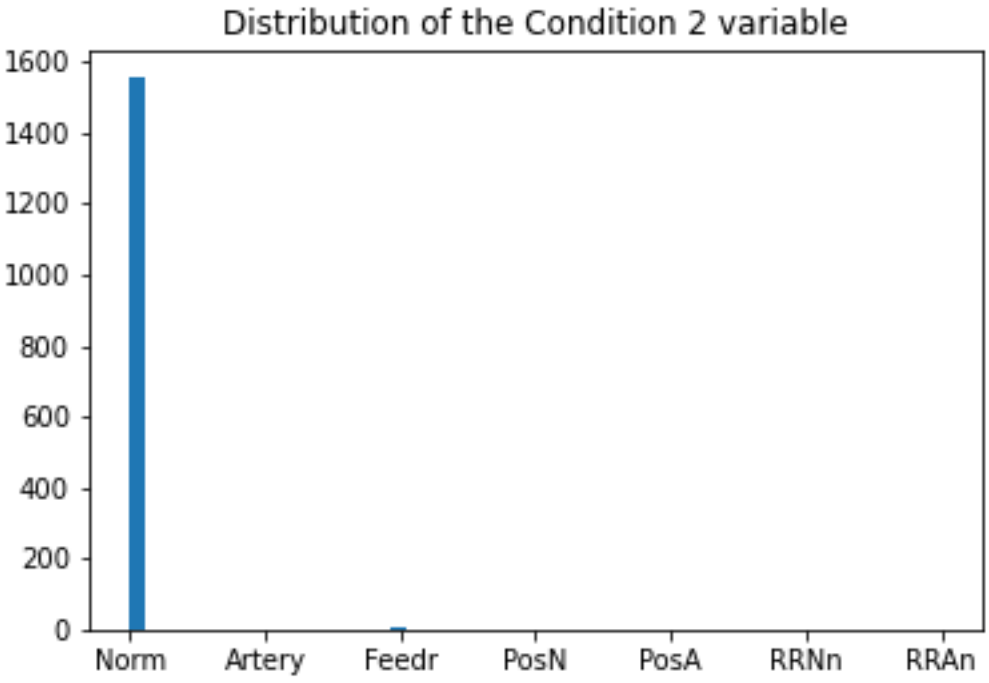

Distribution of the Exterior 1st variable
500
400
300
200
100
0

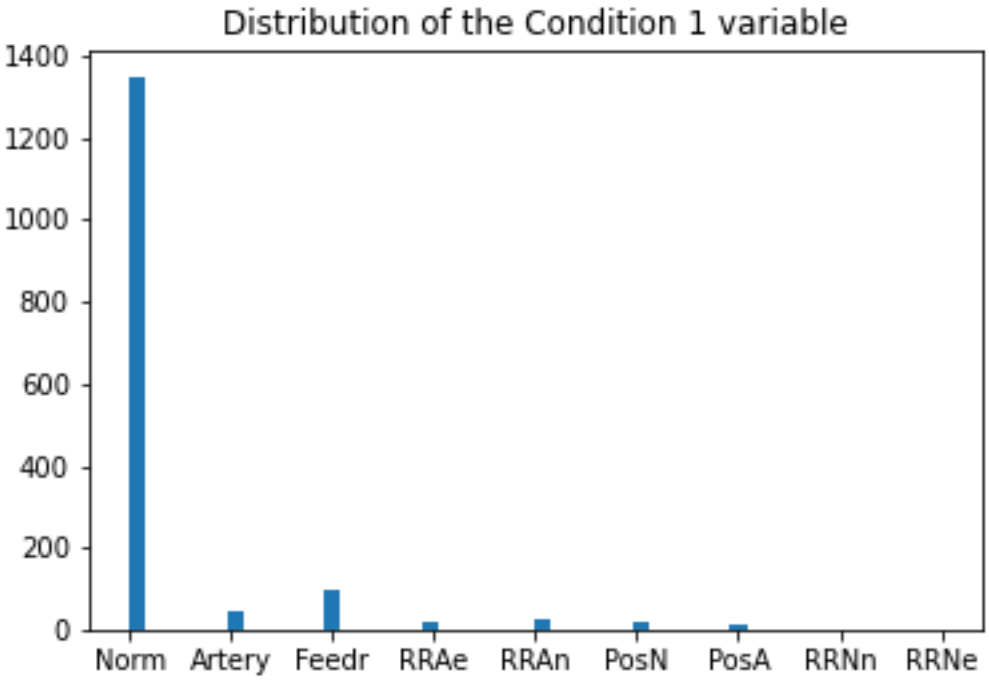

Distribution of the Exter Qual variable
1000
800
600
400
200
0
Gd
TA
Ex
Fa

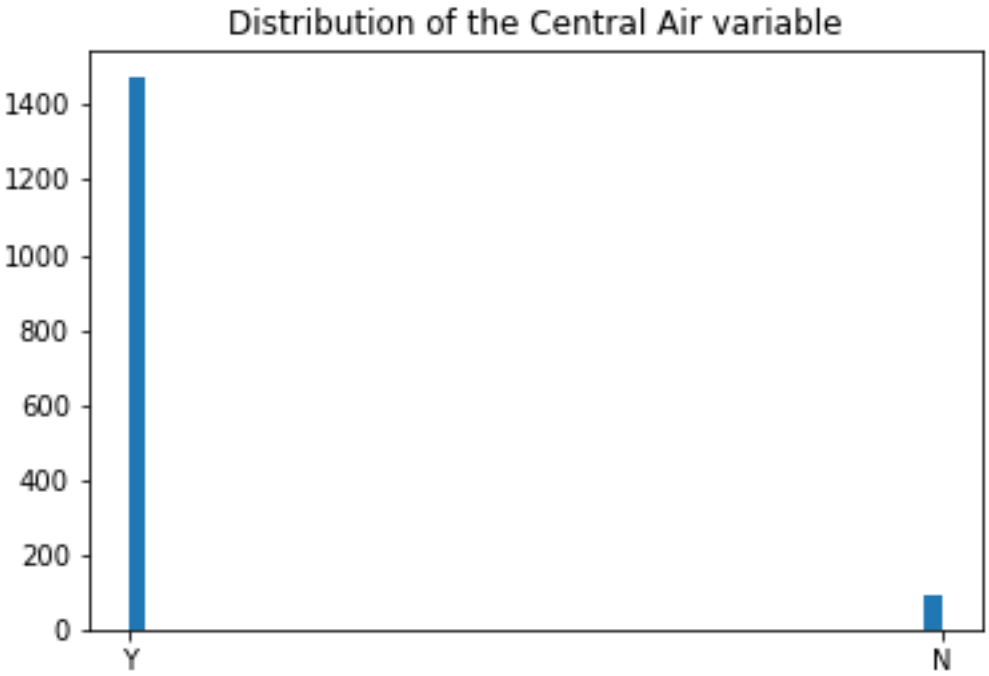

Distribution of the Exter Cond variable
1400
1200
1000
800
600
400
200
0
TA
Gd
Fa
Po
Ex

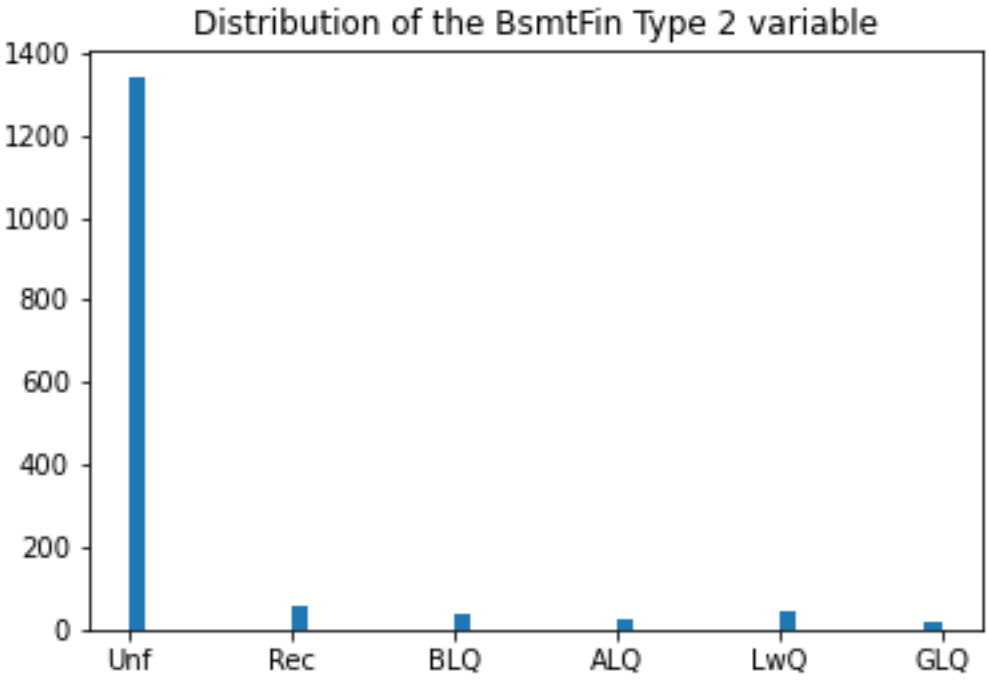

Distribution of the Enclosed Porch variable
1200
1000
800
600
400
200
0
0
100
200
300
400
500
600

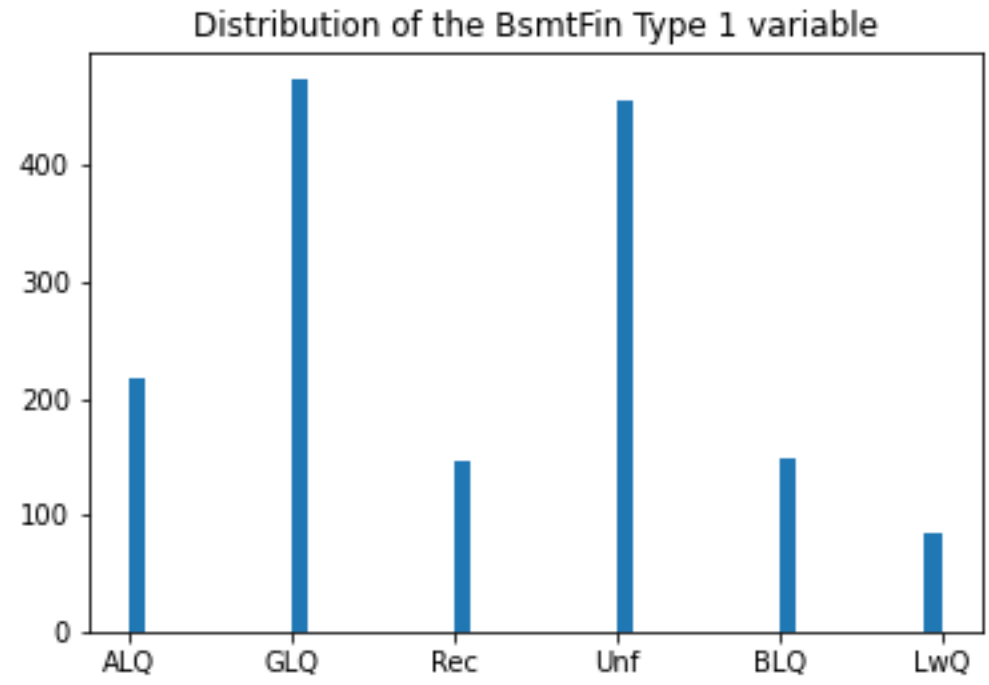

Distribution of the Electrical variable
1400
1200
1000
800
600
400
200
0
SBrkr
FuseA
FuseF
FuseP

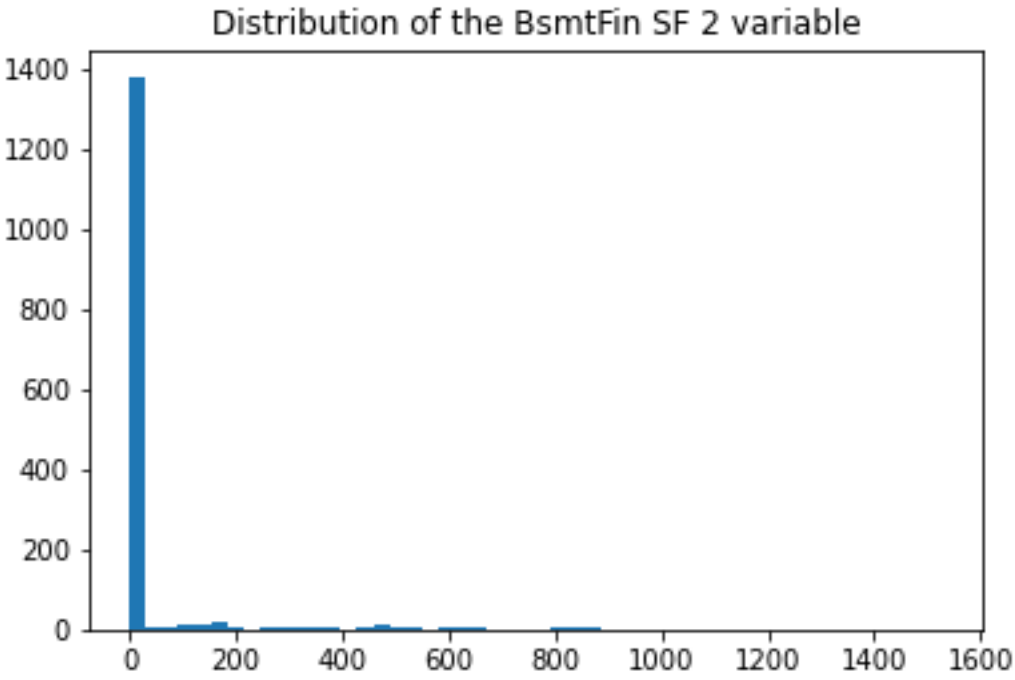

Distribution of the Condition 2 variable
1600
1400
1200
1000
800
600
400
200
0
Norm
Artery
Feedr
PosN
PosA
RRNn
RRAn

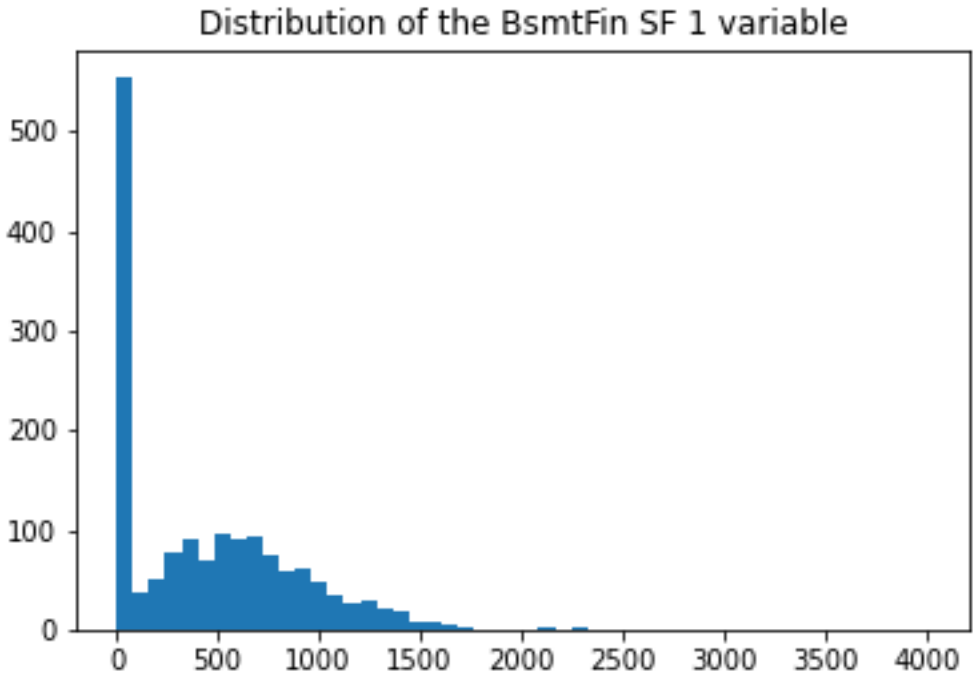

Distribution of the Condition 1 variable
1400
1200
1000
800
600
400
200
0
Norm
Artery
Feedr
RRAe
RRAn
PosN
PosA
RRNn
RRNe

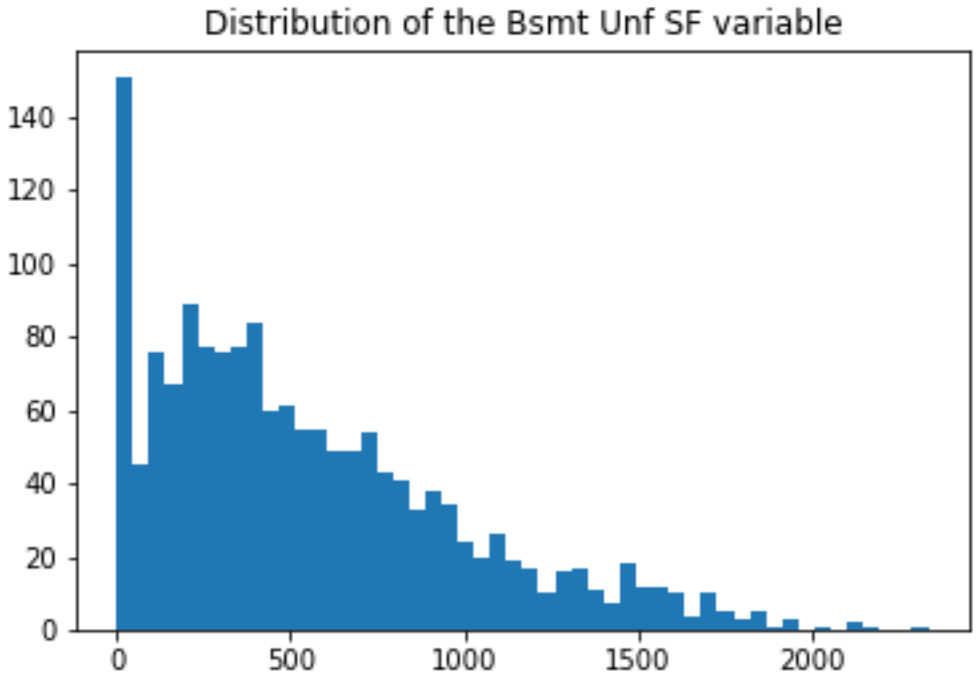

Distribution of the Central Air variable
1400
1200
1000
800
600
400
200
0
Y
N

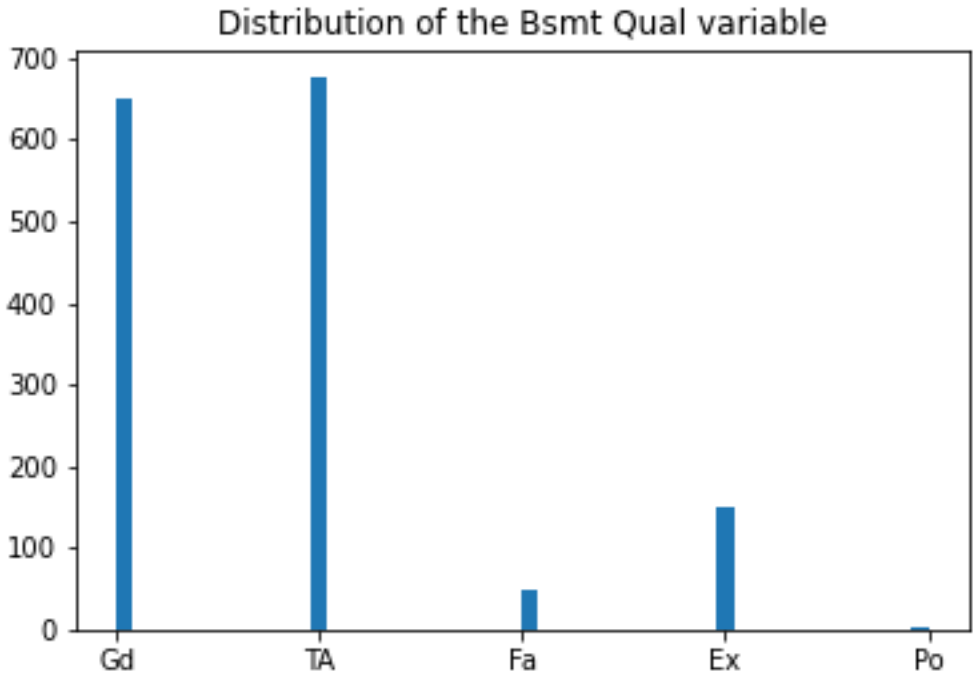

Distribution of the BsmtFin Type 2 variable
1400
1200
1000
800
600
400
200
0
Unf
Rec
BLQ
ALQ
LwQ
GLQ

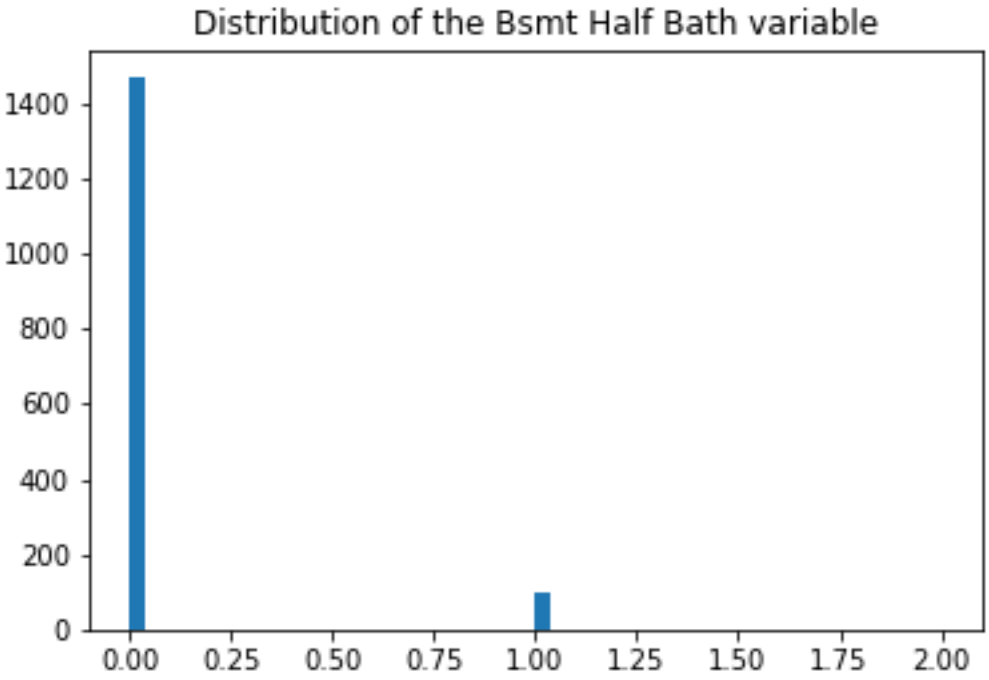

Distribution of the BsmtFin Type 1 variable
400
300
200
100
0
ALQ
GLQ
Rec
Unf
BLQ
LwQ

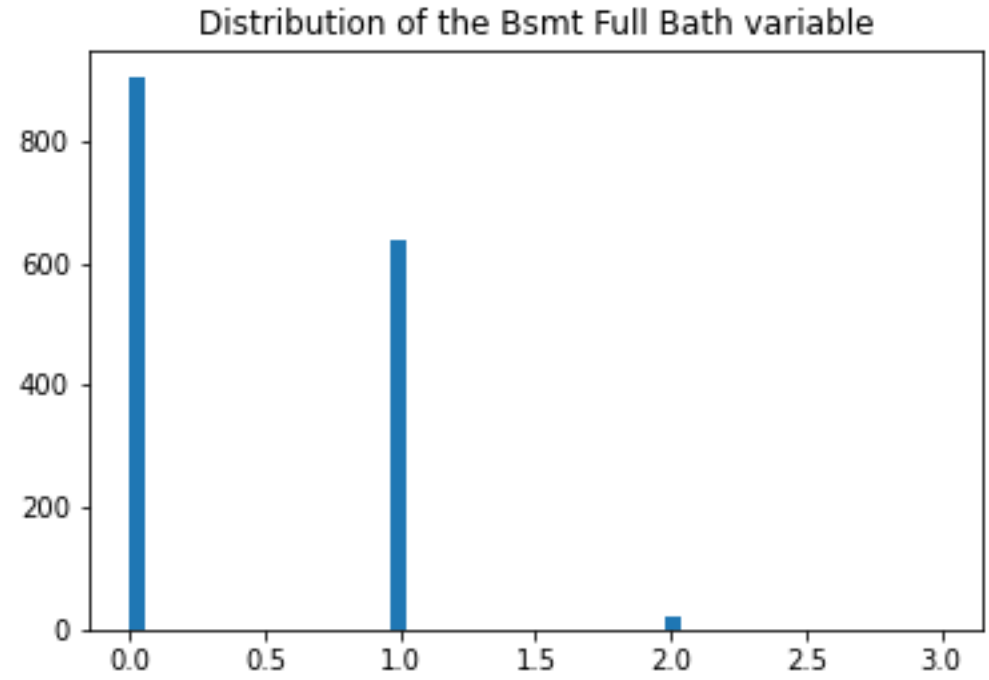

Distribution of the BsmtFin SF 2 variable
1400
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
1200
1400
1600

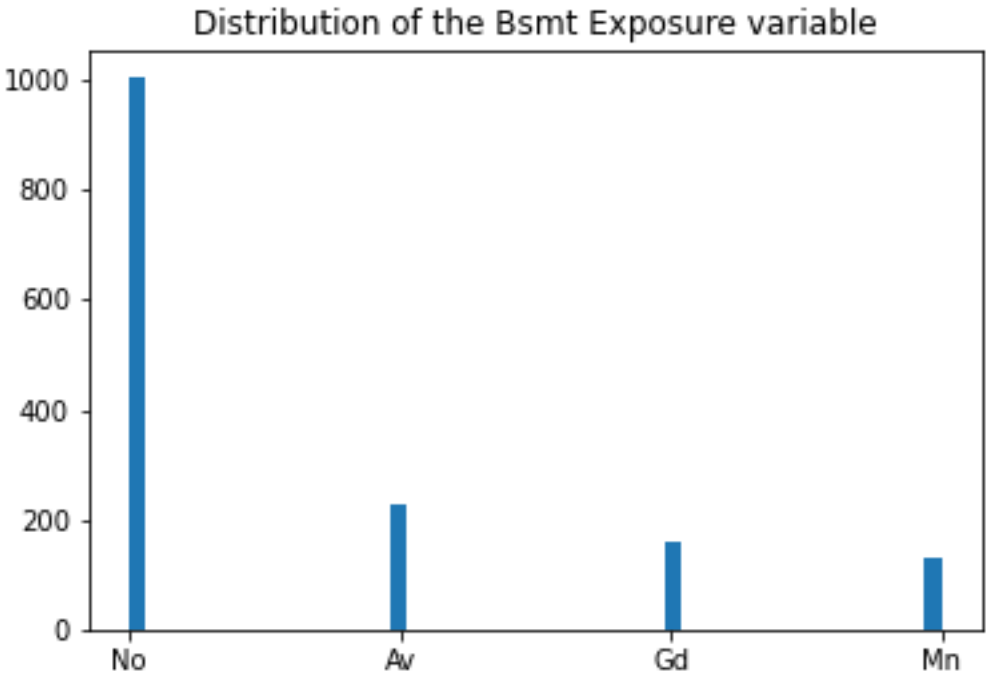

Distribution of the BsmtFin SF 1 variable
500
400
300
200
100
0
0
500
1000
1500
2000
2500
3000
3500
4000

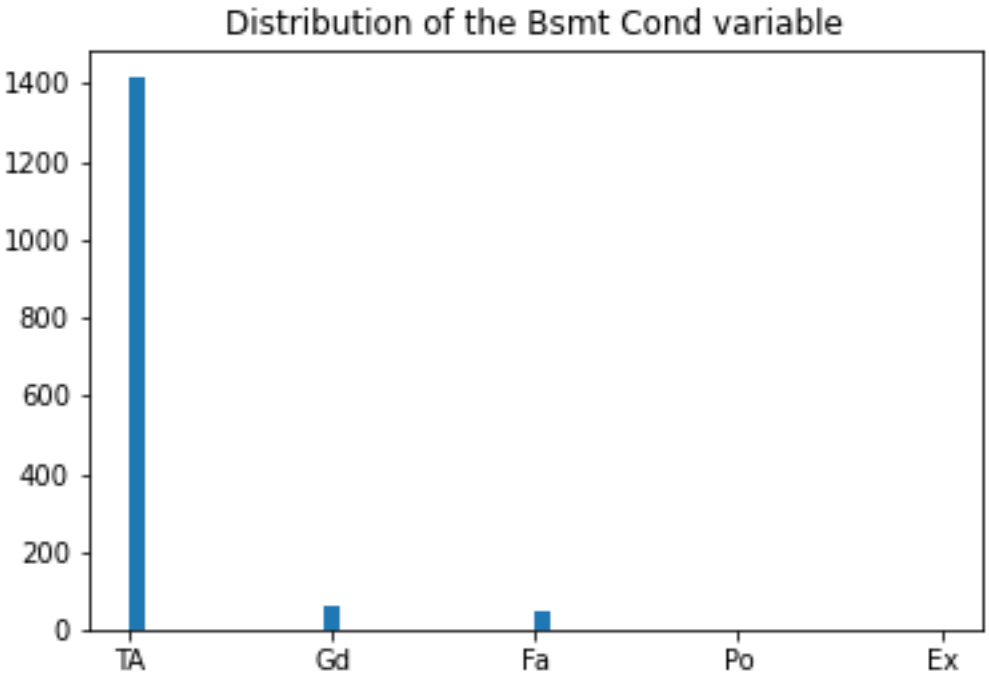

Distribution of the Bsmt Unf SF variable
140
120
100
80
60
40
20
0
0
500
1000
1500
2000

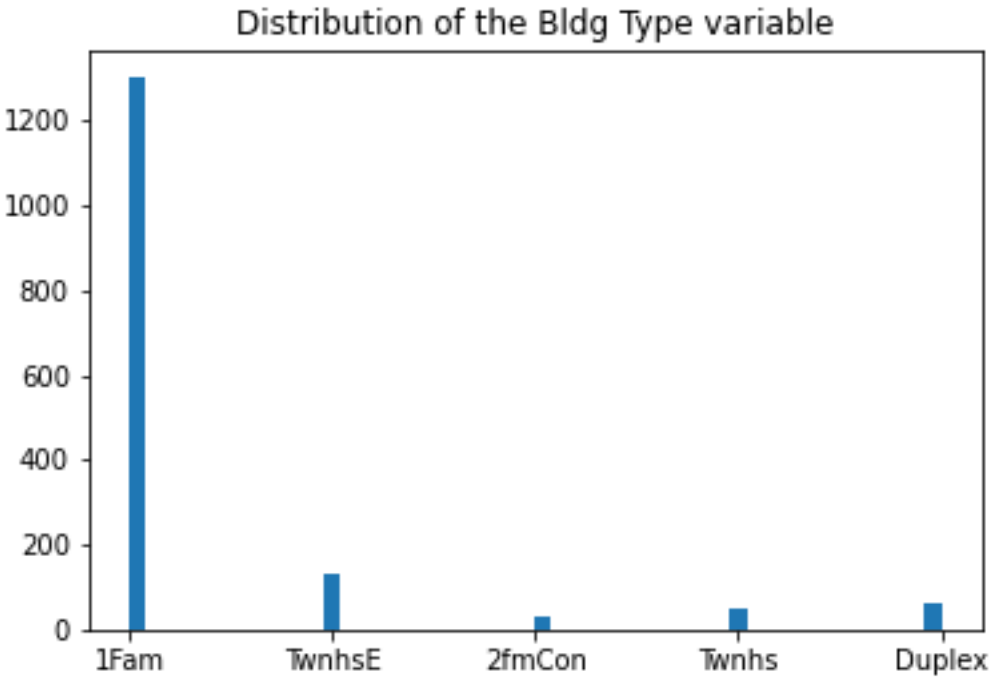

Distribution of the Bsmt Qual variable
700
600
500
400
300
200
100
0
Gd
TA
Fa
Ex
Po

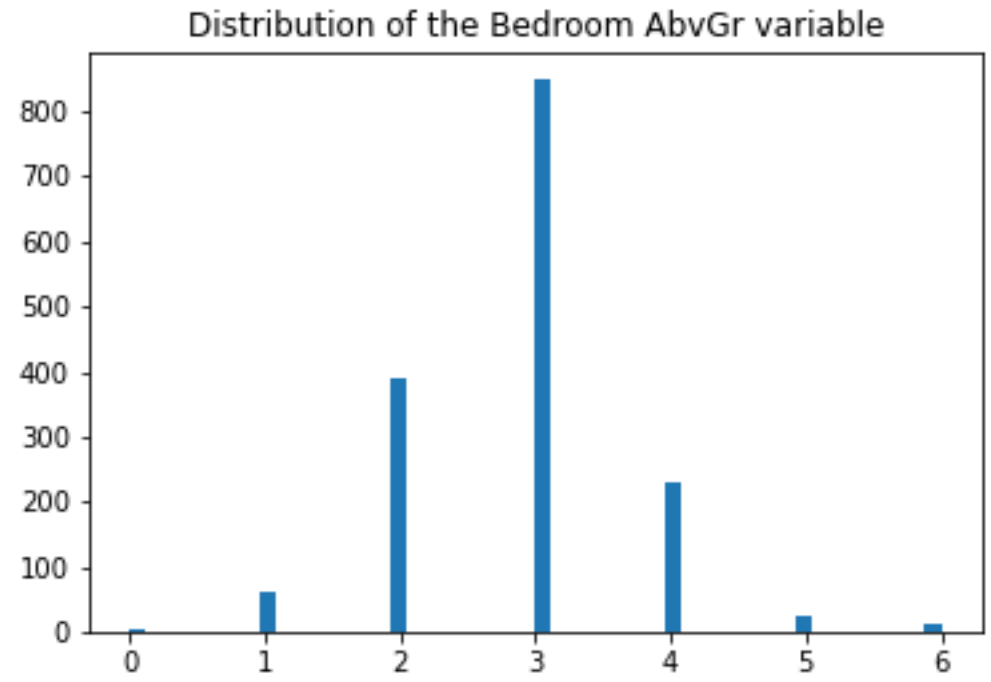

Distribution of the Bsmt Half Bath variable
1400
1200
1000
800
600
400
200
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

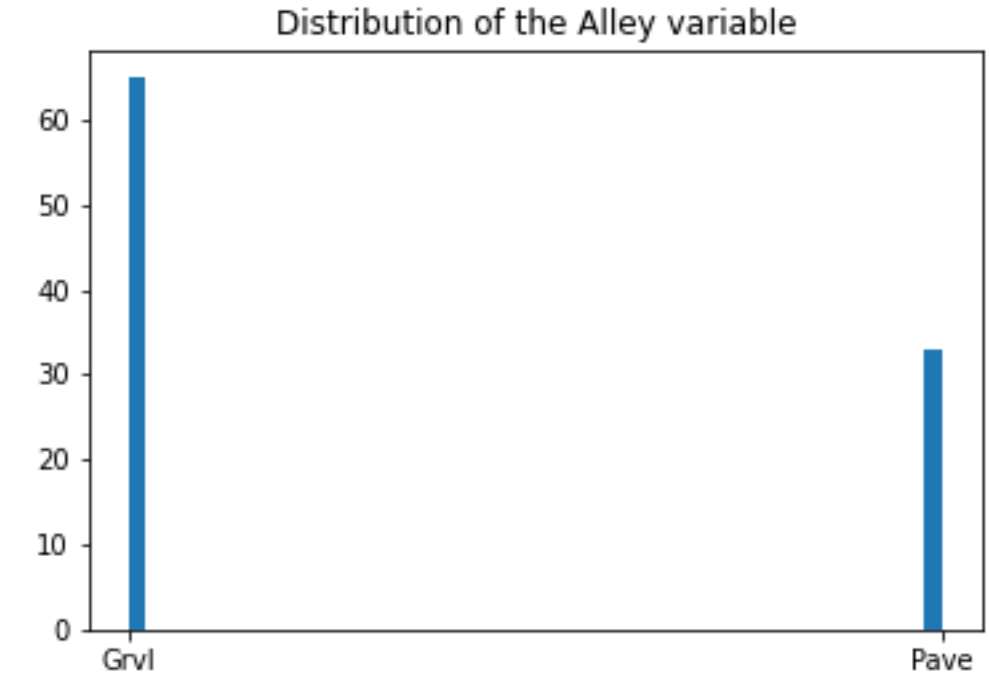

Distribution of the Bsmt Full Bath variable
800
600
400
200
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0

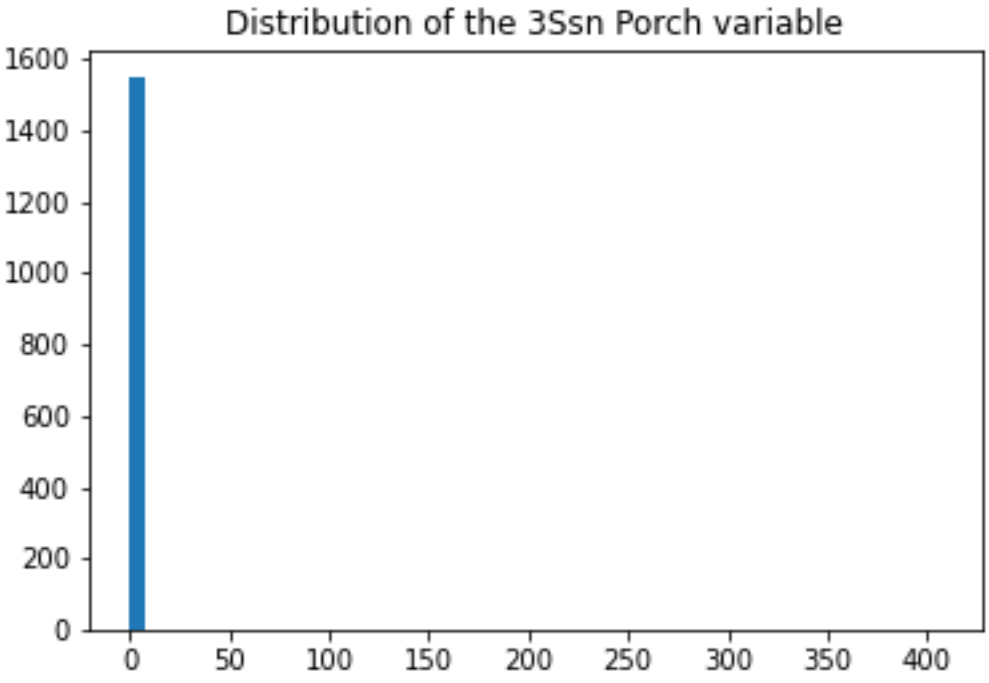

Distribution of the Bsmt Exposure variable
1000
800
600
400
200
0
No
Av
Gd
Mn

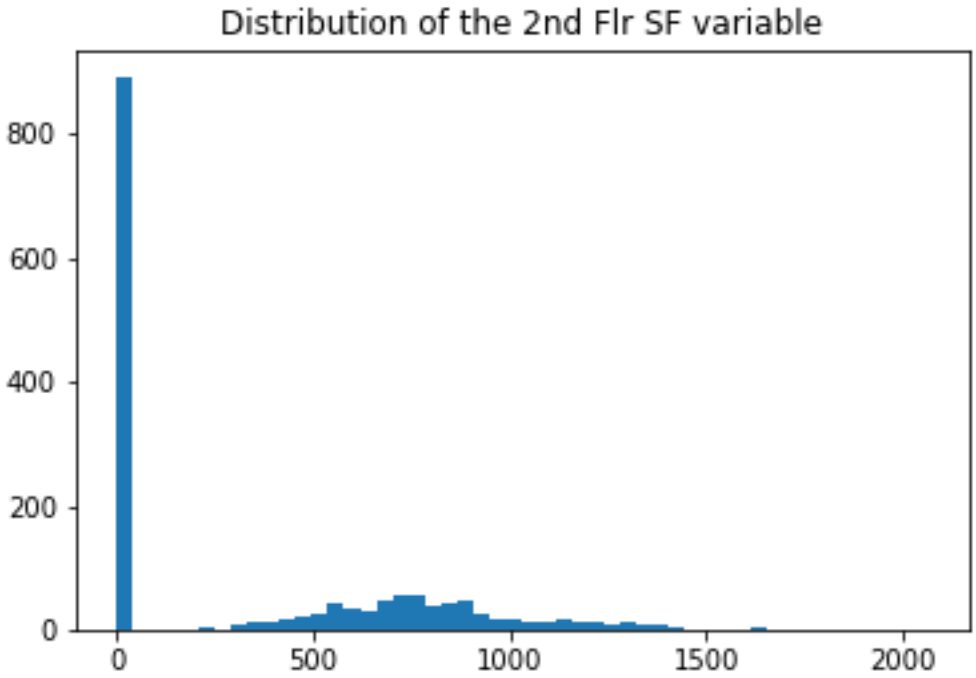

Distribution of the Bsmt Cond variable
1400
1200
1000
800
600
400
200
0
TA
Gd
Fa
Po
Ex

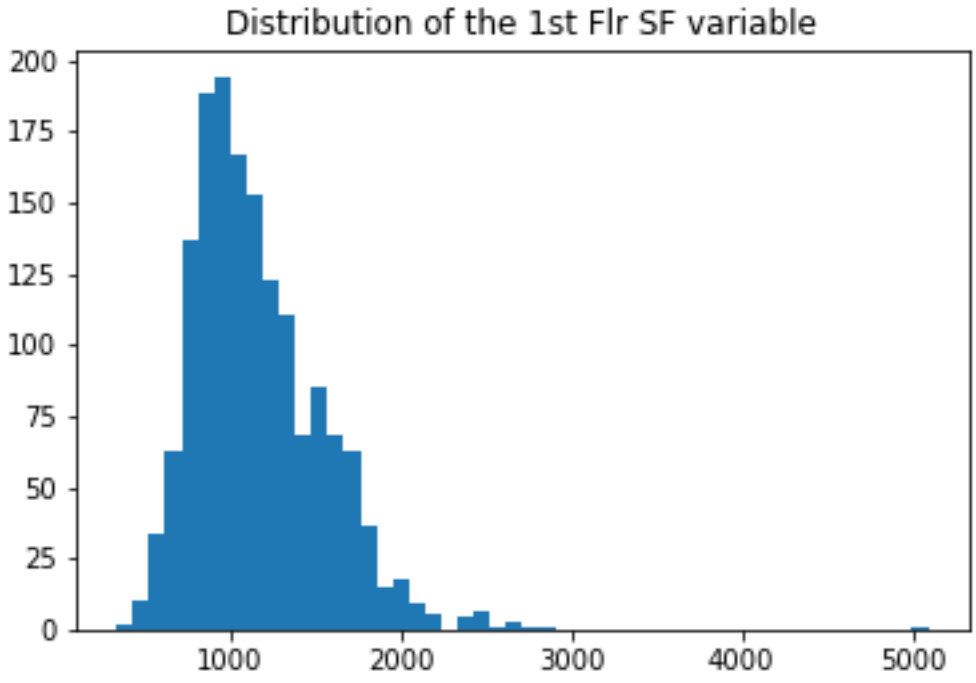

Distribution of the Bldg Type variable
1200
1000
800
600
400
200
0
1Fam
TwnhsE
2fmCon
Twnhs
Duplex

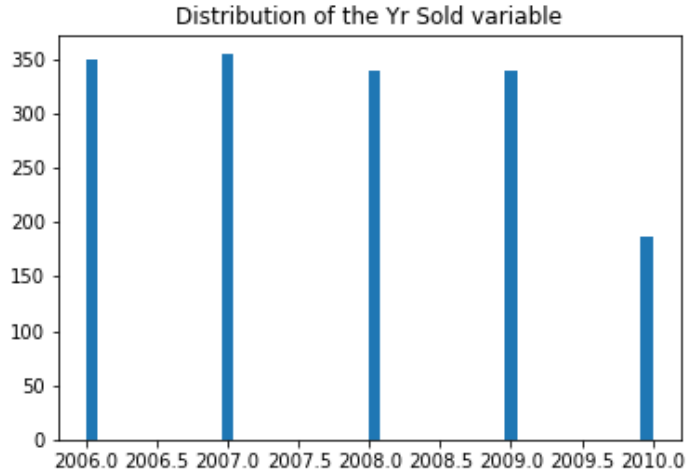

Distribution of the Bedroom AbvGr variable
800
700
600
500
400
300
200
100
0
0
1
2
3
4
5
6


Distribution of the Alley variable
60
50
40
30
20
10
0
Grvl
Pave

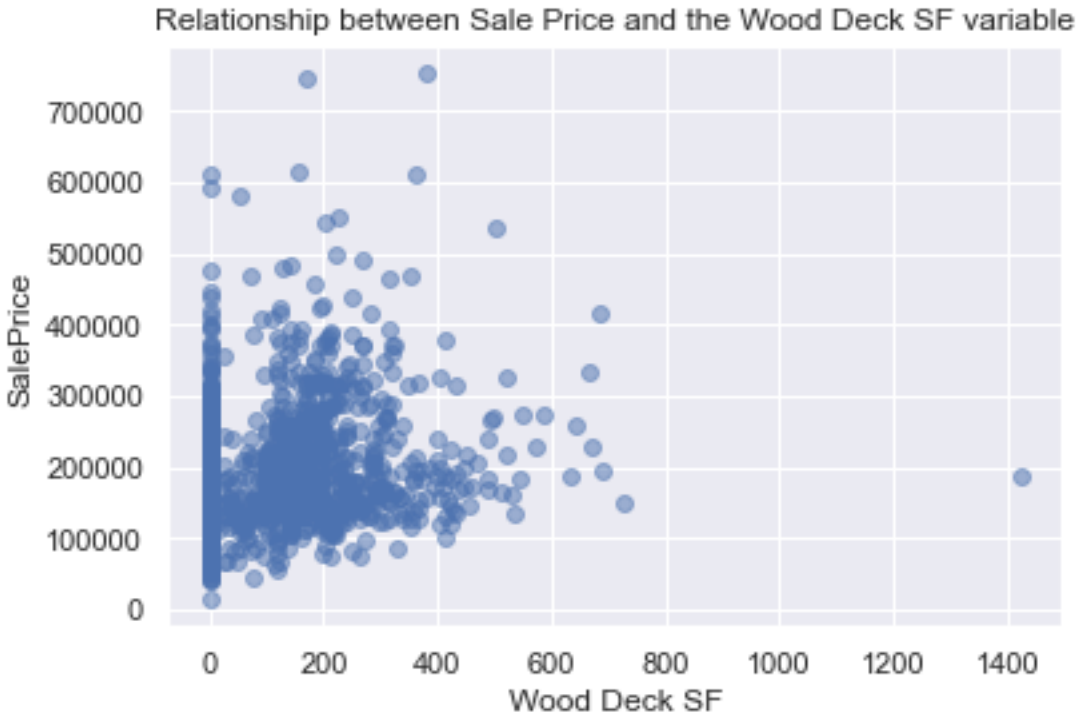

Distribution of the 3Ssn Porch variable
1600
1400
1200
1000
800
600
400
200
0
0
50
100
150
200
250
300
350
400

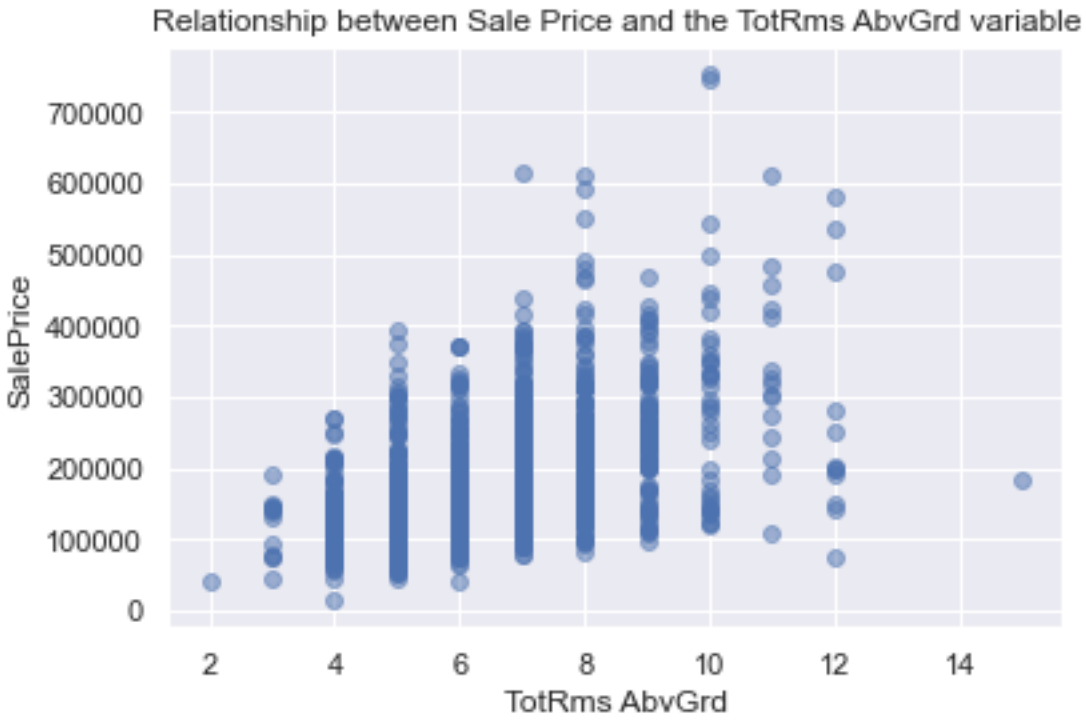

Distribution of the 2nd Flr SF variable
800
600
400
200
0
0
500
1000
1500
2000

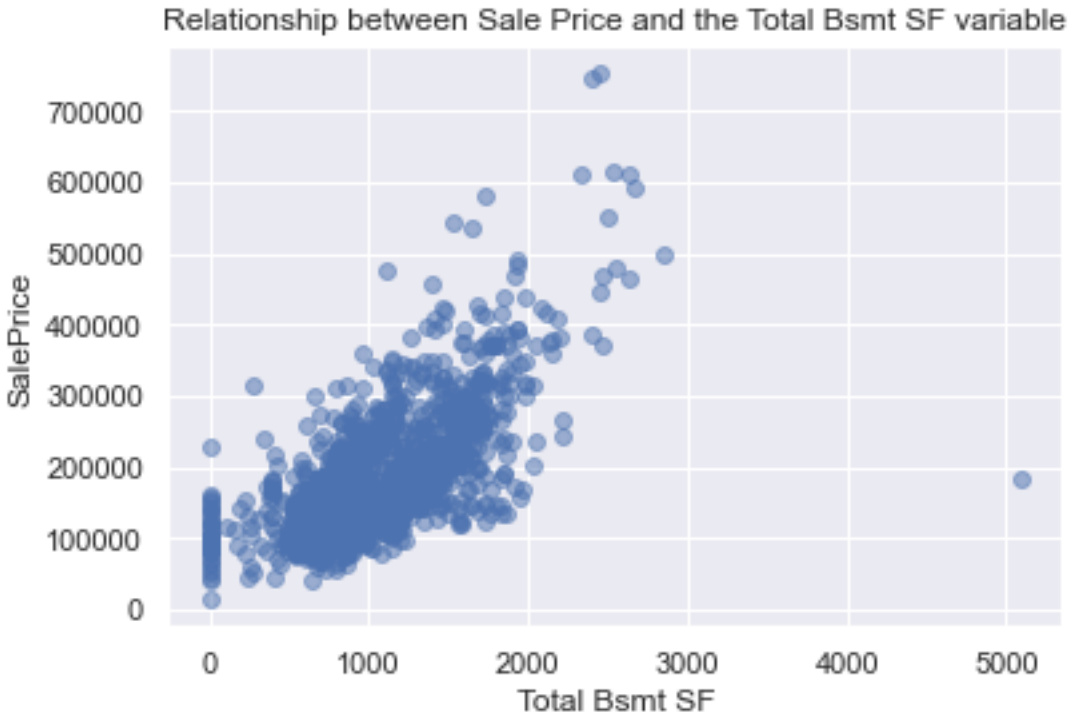

Distribution of the 1st Flr SF variable
200
175
150
125
100
75
50
25
0
1000
2000
3000
4000
5000



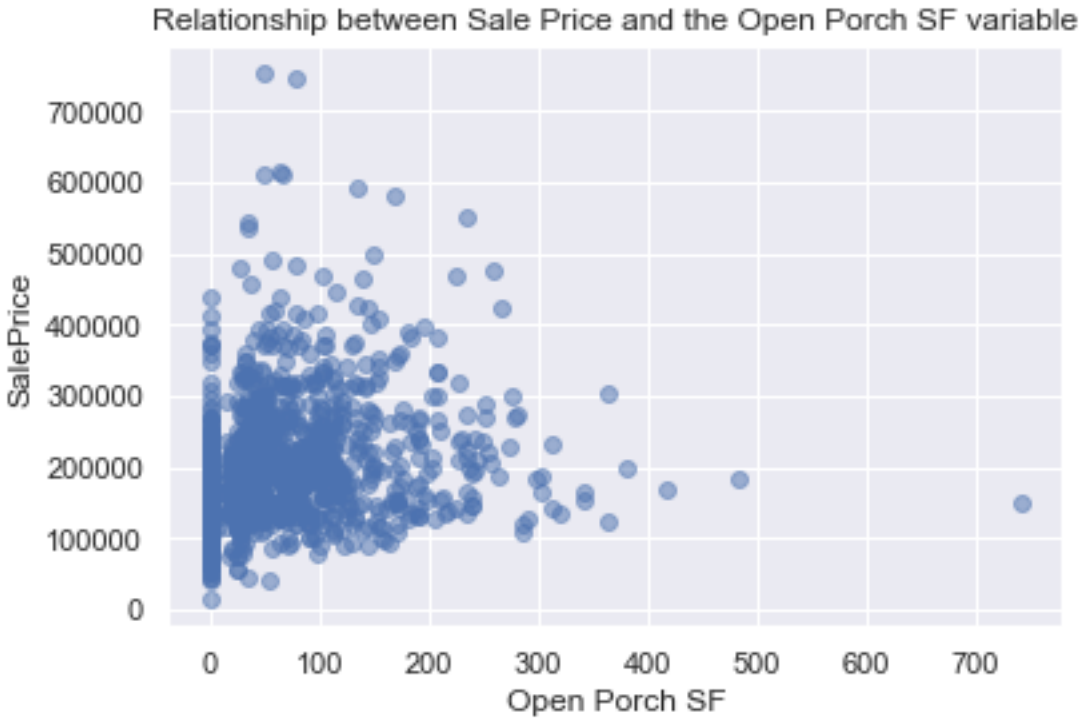
Appendix iii



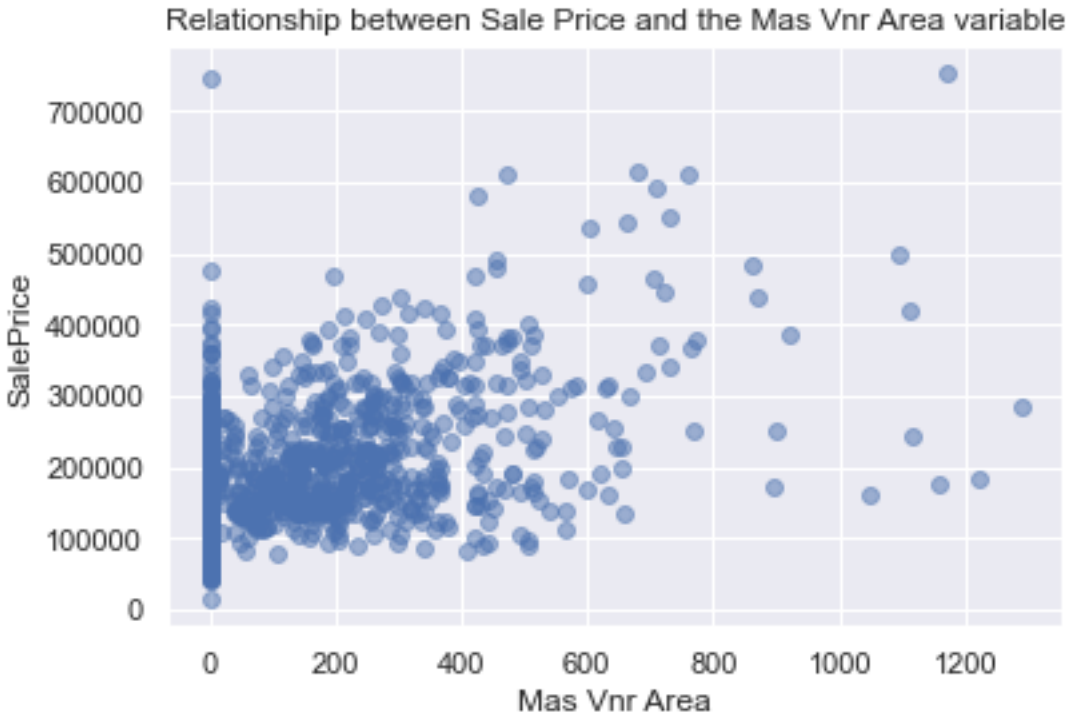

Relationship between Sale Price and the TotRms AbvGrd variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
2
4
6
8
10
12
14
TotRms AbvGrd


Relationship between Sale Price and the Total Bsmt SF variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
0
1000
2000
3000
4000
5000
Total Bsmt SF

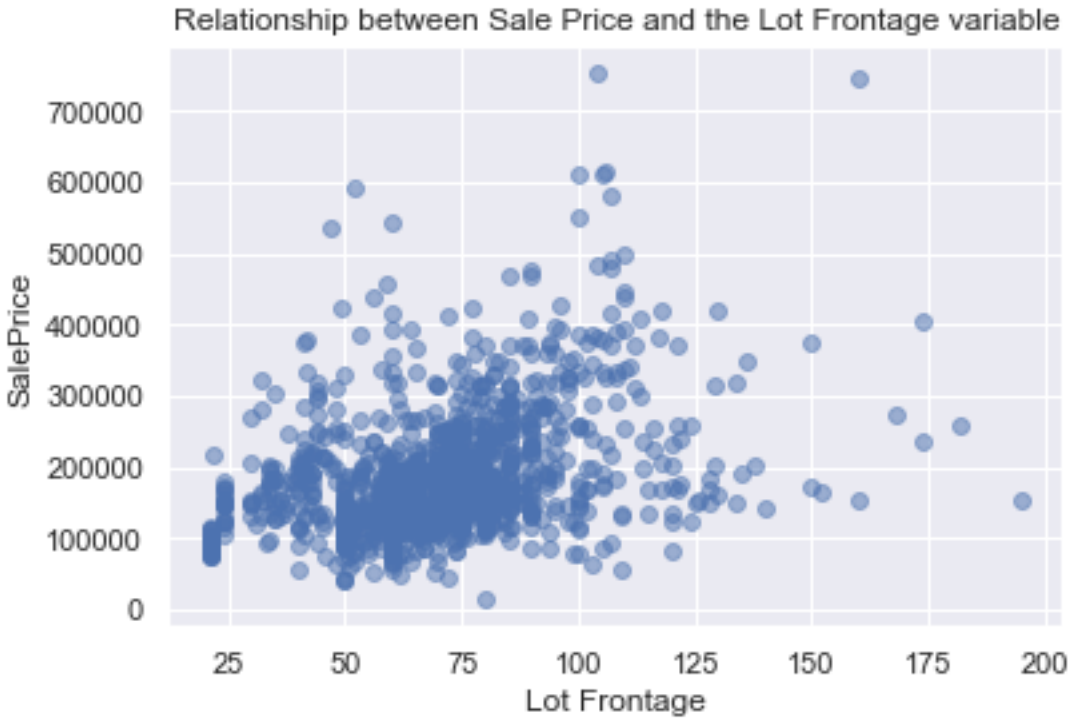

Relationship between Sale Price and the Screen Porch variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
0
100
200
300
400
500
Screen Porch


Relationship between Sale Price and the Open Porch SF variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
0
100
200
300
400
500
600
700
Open Porch SF


Relationship between Sale Price and the Misc Val variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
0
2500
5000
7500
10000
12500
15000
17500
Misc Val

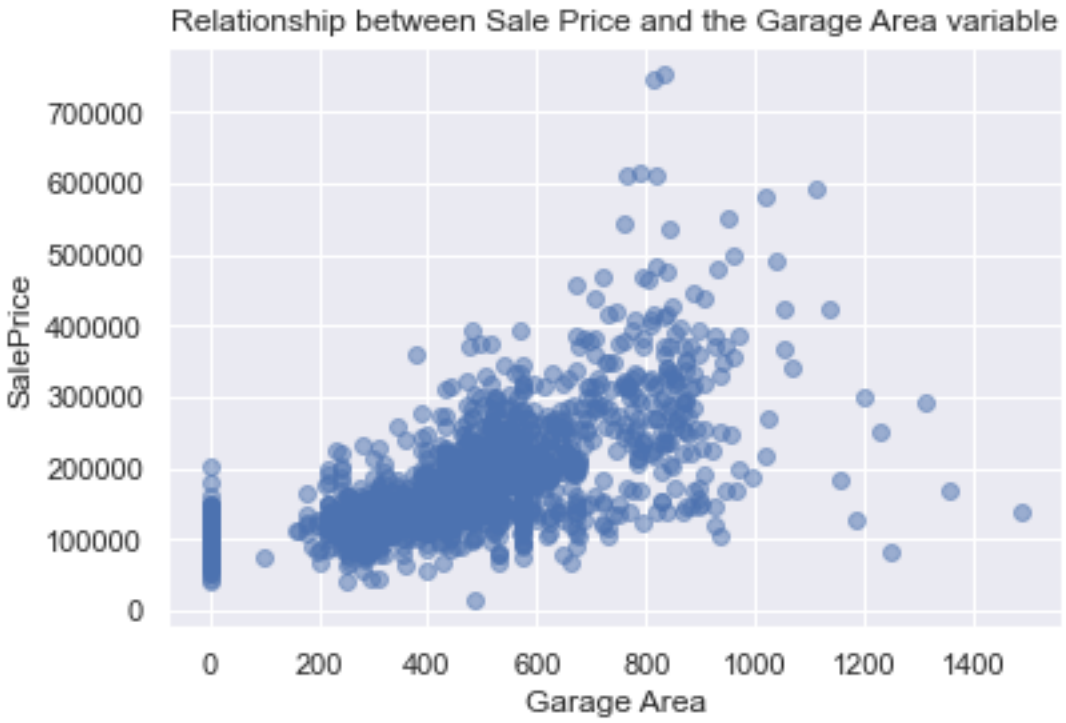

Relationship between Sale Price and the Mas Vnr Area variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
0
200
400
600
800
1000
1200
Mas Vnr Area


Relationship between Sale Price and the Low Qual Fin SF variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
0
200
400
600
800
1000
Low Qual Fin SF

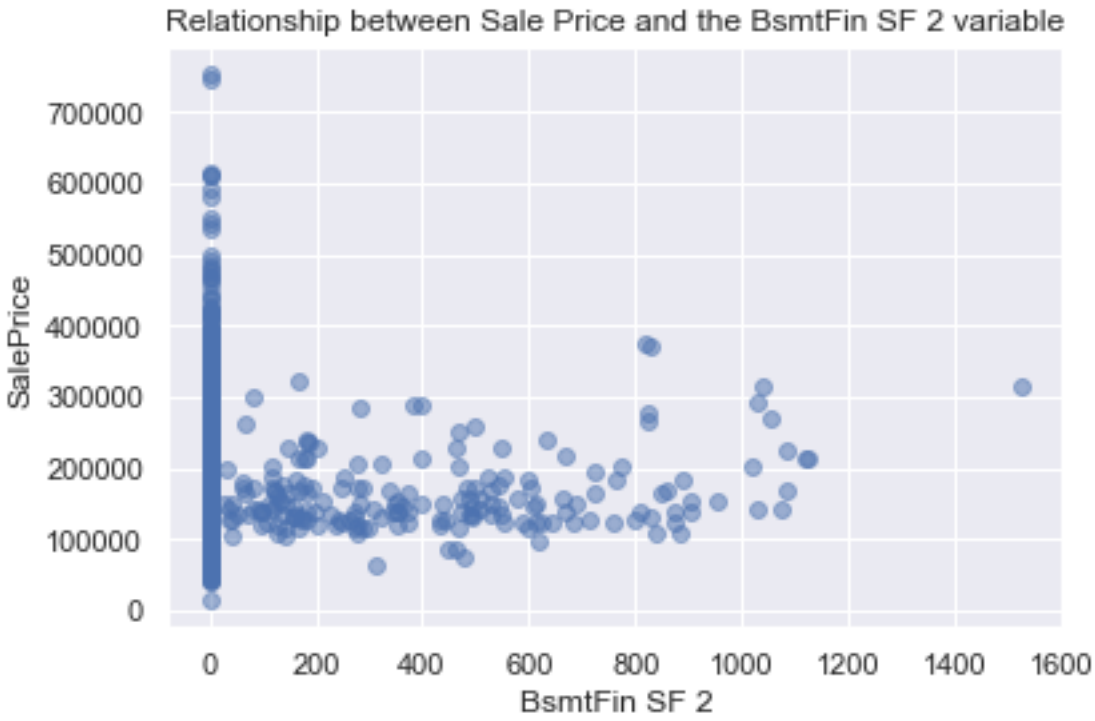

Relationship between Sale Price and the Lot Frontage variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
25
50
75
100
125
150
175
200
Lot Frontage

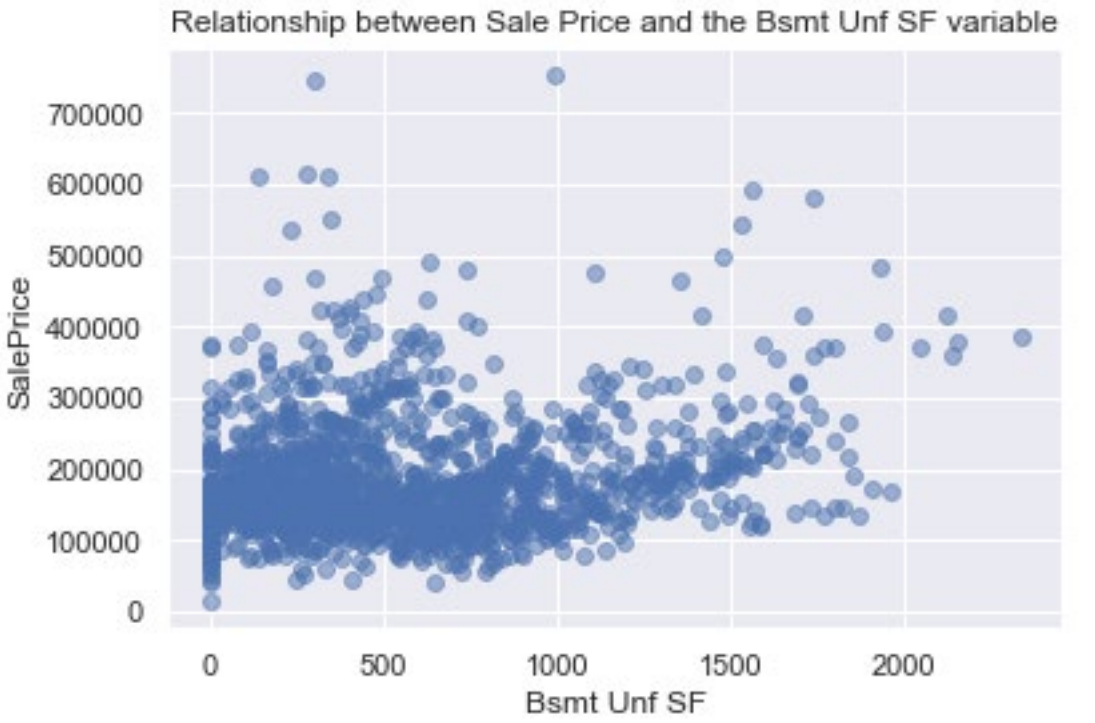

Relationship between Sale Price and the Lot Area variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
0
50000
100000
150000
200000
Lot Area


Relationship between Sale Price and the Last remod/add variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
0
10
20
30
40
50
60
Last remod/add


Relationship between Sale Price and the Garage Area variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
0
200
400
600
800
1000
1200
1400
Garage Area

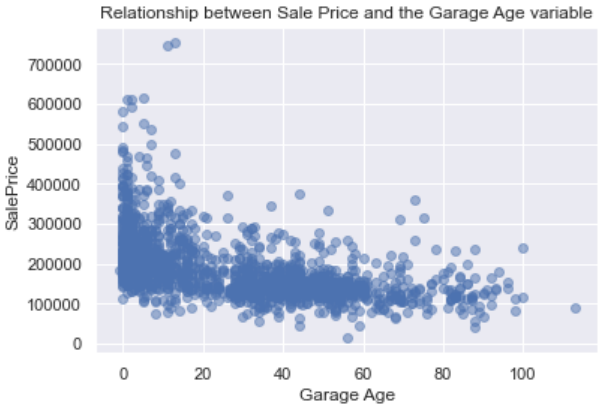

Relationship between Sale Price and the Enclosed Porch variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
0
100
200
300
400
500
600
Enclosed Porch


Relationship between Sale Price and the BsmtFin SF 2 variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
0
200
400
600
800
1000
1200
1400
1600
BsmtFin SF 2

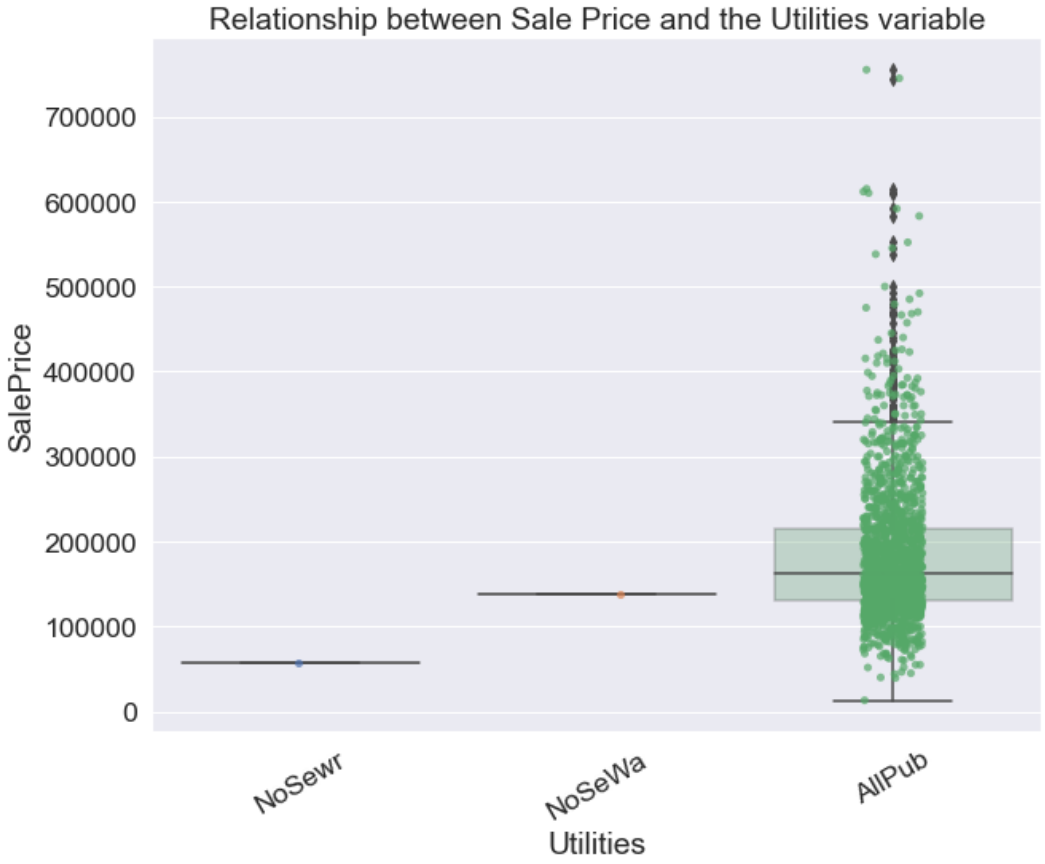

Relationship between Sale Price and the Bsmt Unf SF variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
0
500
1000
1500
2000
Bsmt Unf SF


Relationship between Sale Price and the 3Ssn Porch variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
0
50
100
150
200
250
300
350
400
3Ssn Porch

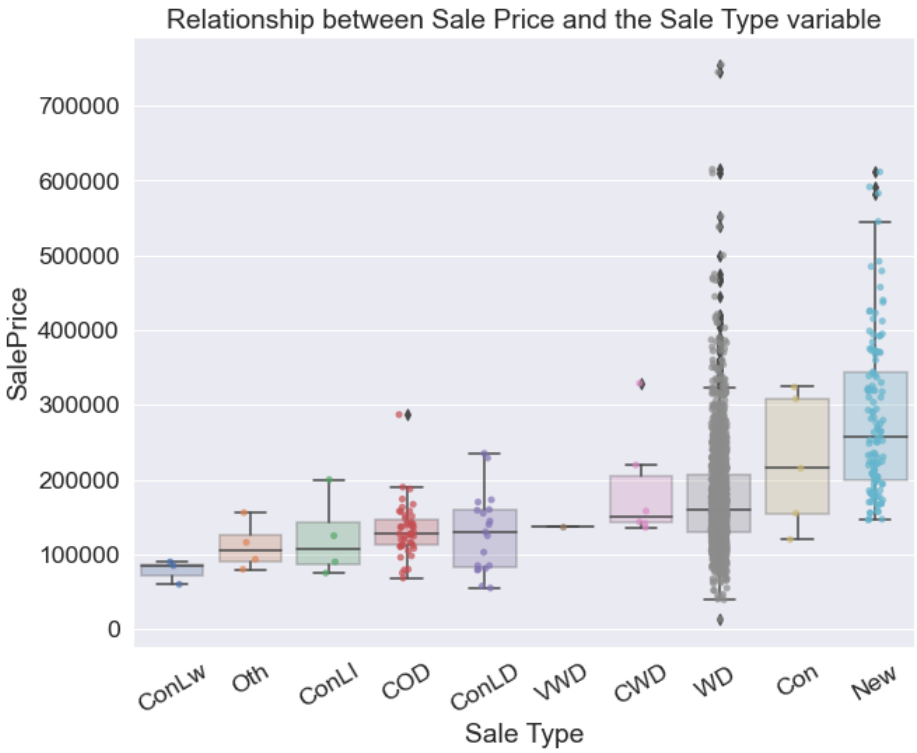

Relationship between Sale Price and the 2nd Flr SF variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
0
500
1000
1500
2000
2nd Flr SF

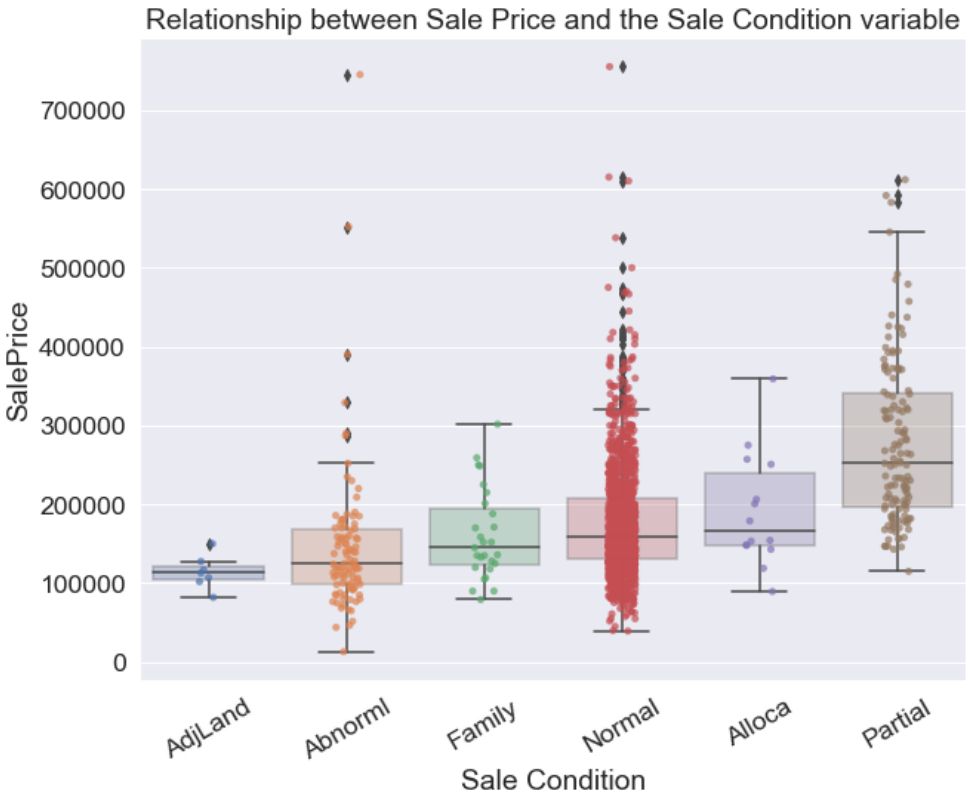

Relationship between Sale Price and the Garage Age variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
0
20
40
60
80
100
Garage Age

Appendix iv

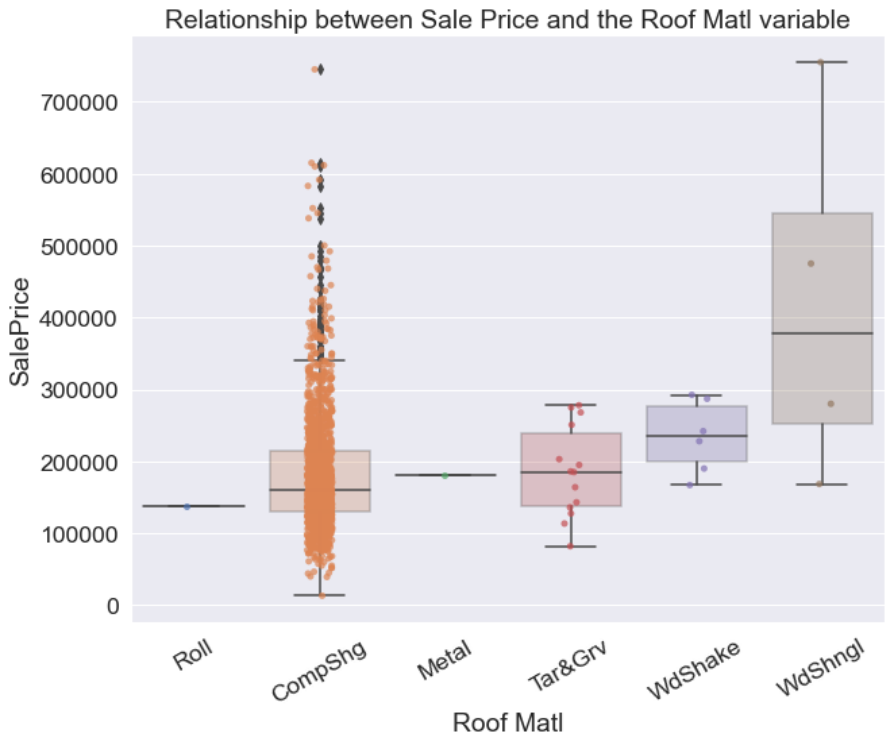

Relationship between Sale Price and the Utilities variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
NoSewr
NoSeWa
AllPub
Utilities

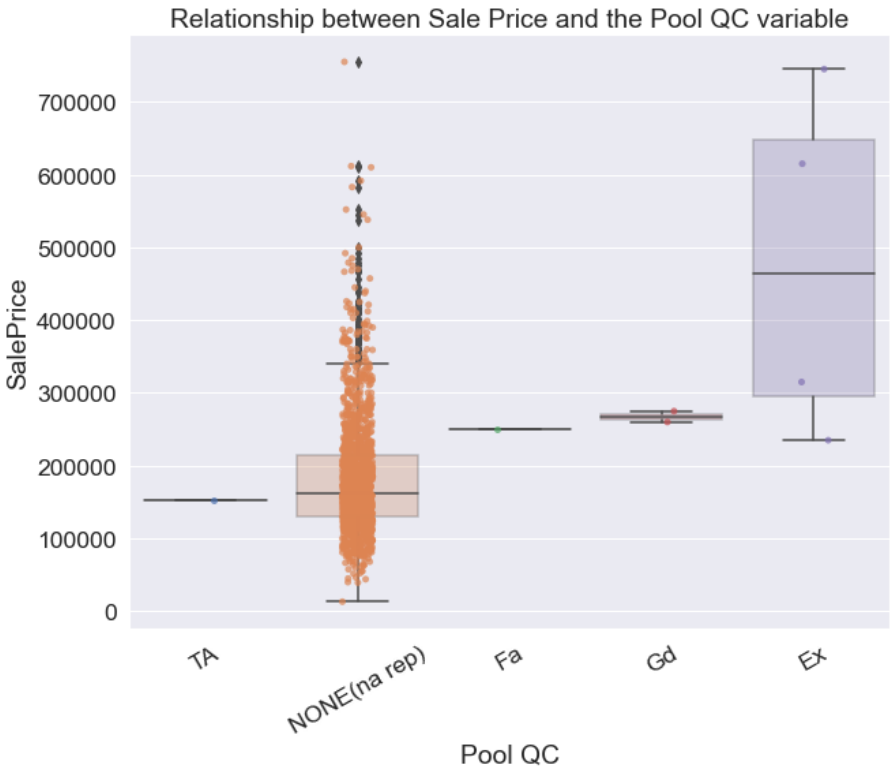

Relationship between Sale Price and the Street variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Grvl
Pave
Street

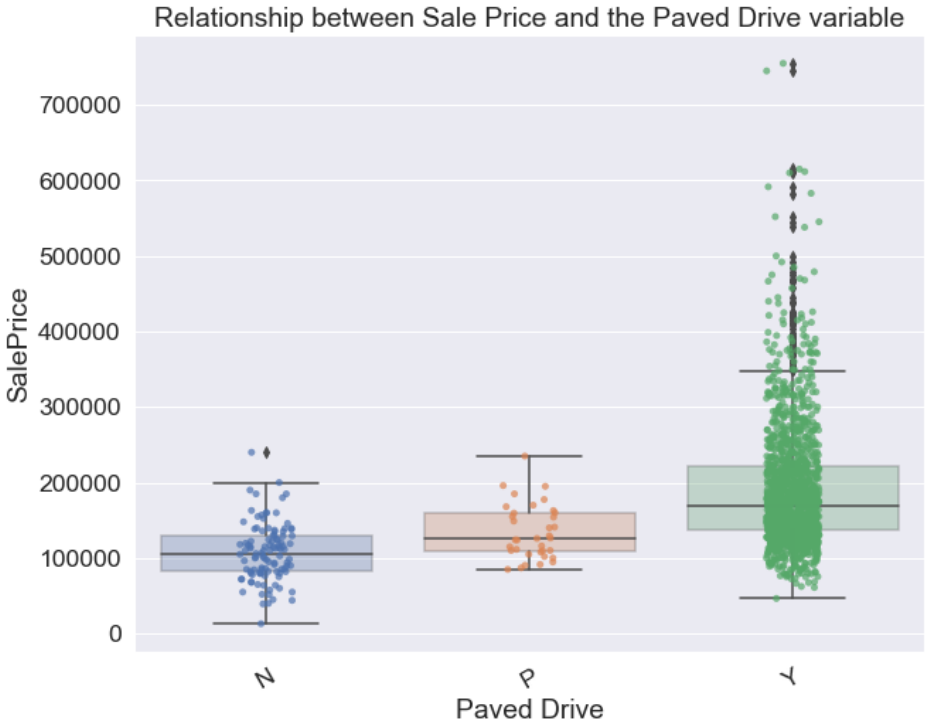

Relationship between Sale Price and the Sale Type variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
ConLw
Oth
ConLI
COD
ConLD
VWD
CWD
WD
Con
New
Sale Type

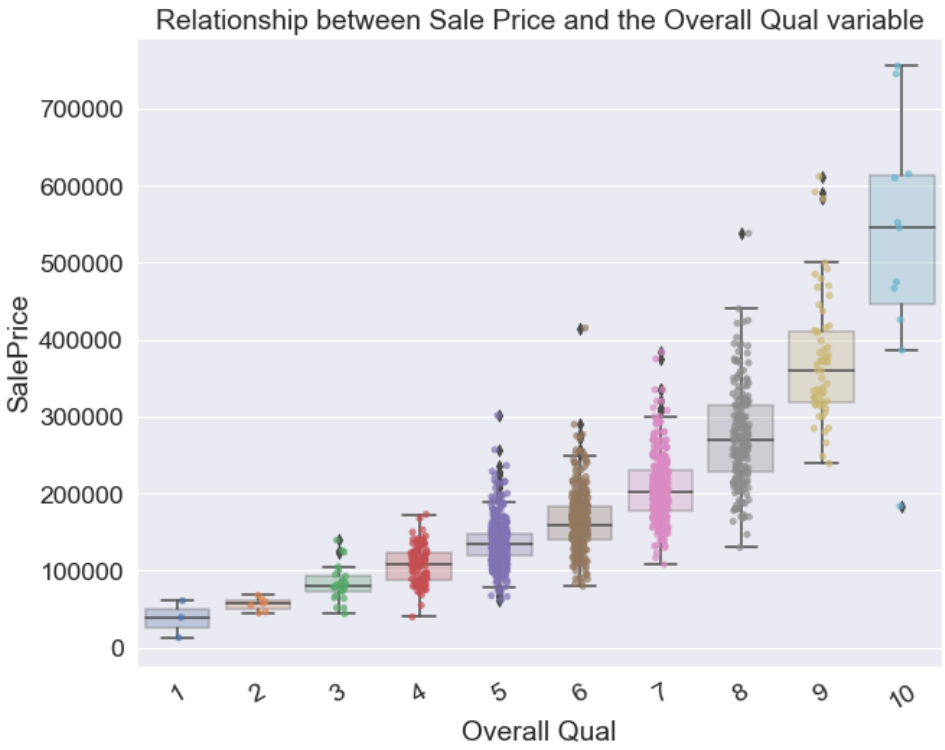

Relationship between Sale Price and the Sale Condition variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
AdjLand
Abnorml
Family
Normal
Alloca
Partial
Sale Condition

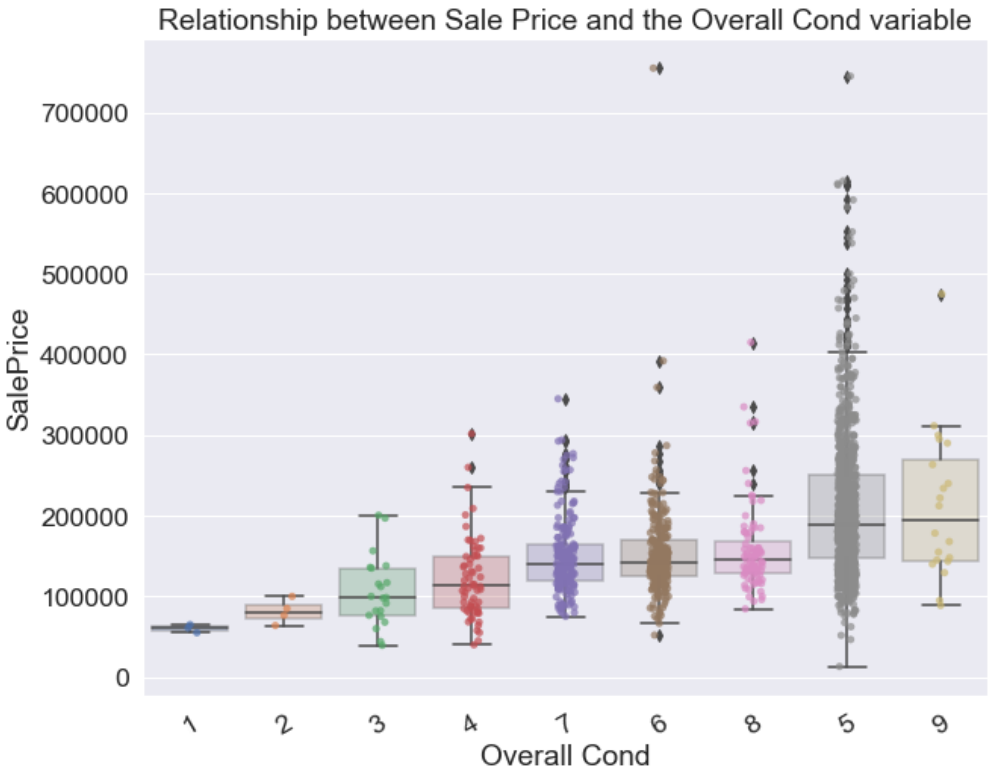

Relationship between Sale Price and the Roof Style variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Gable
Flat
Hip
Mansard
Gambrel
Roof Style

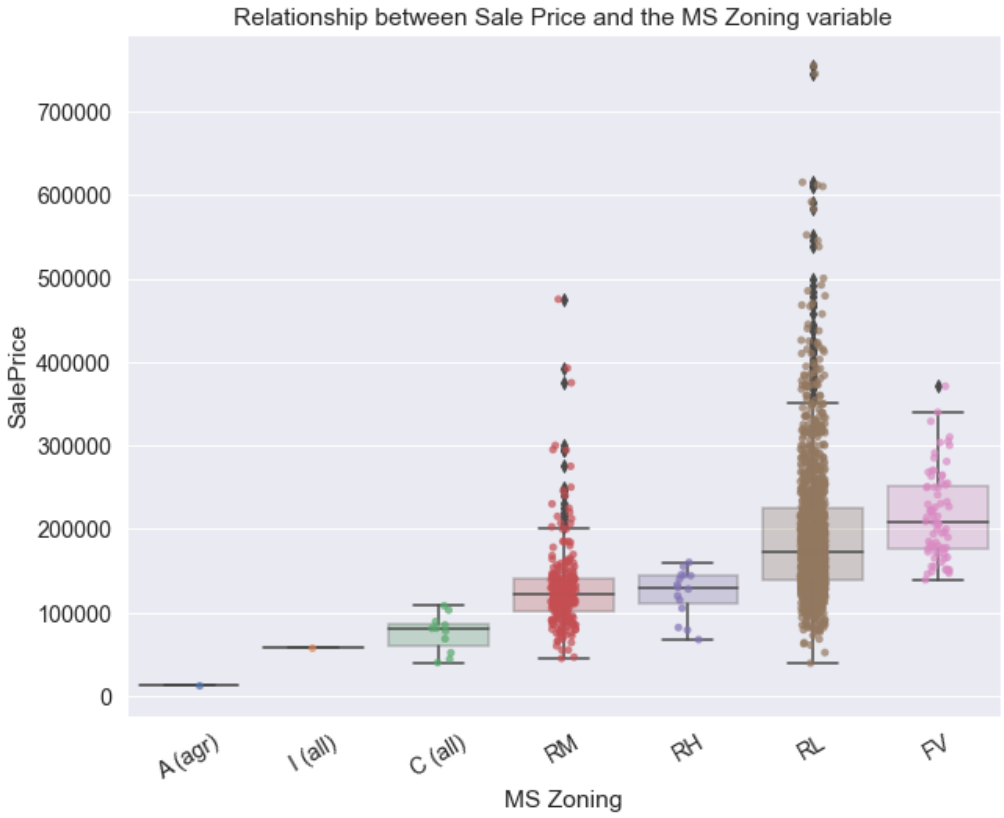

Relationship between Sale Price and the Roof Matl variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Roll
CompShg
Metal
Tar&Grv
WdShake
WdShngl
Roof Matl


Relationship between Sale Price and the Pool QC variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
TA
NONE(na rep)
Fa
Gd
Ex
Pool QC

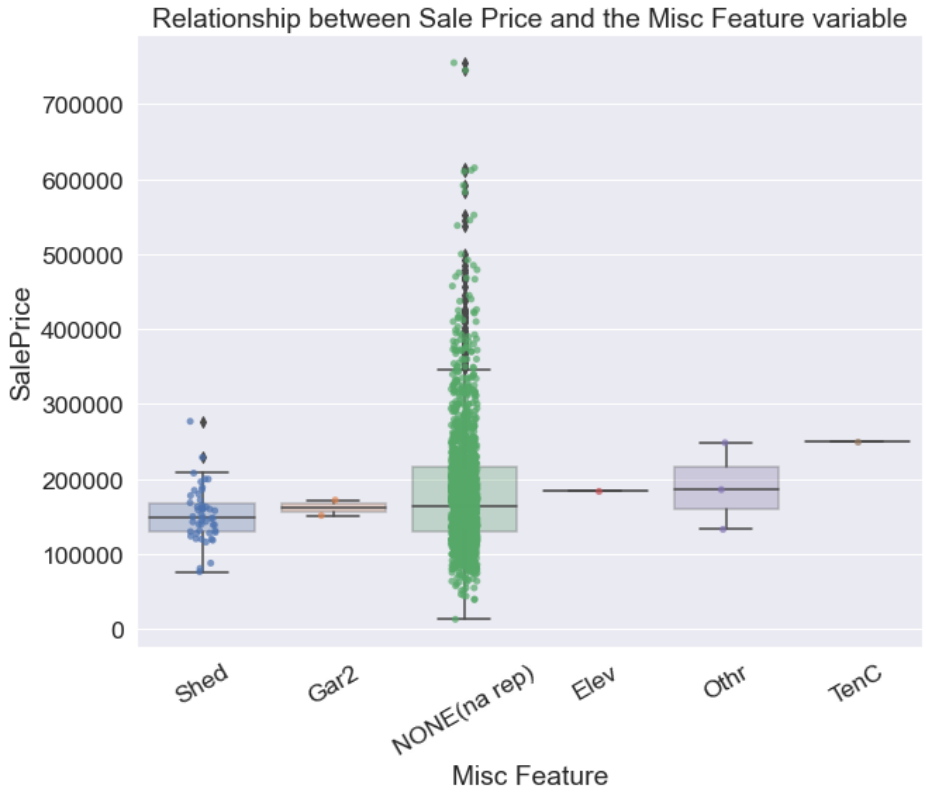

Relationship between Sale Price and the Paved Drive variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
N
P
Y
Paved Drive

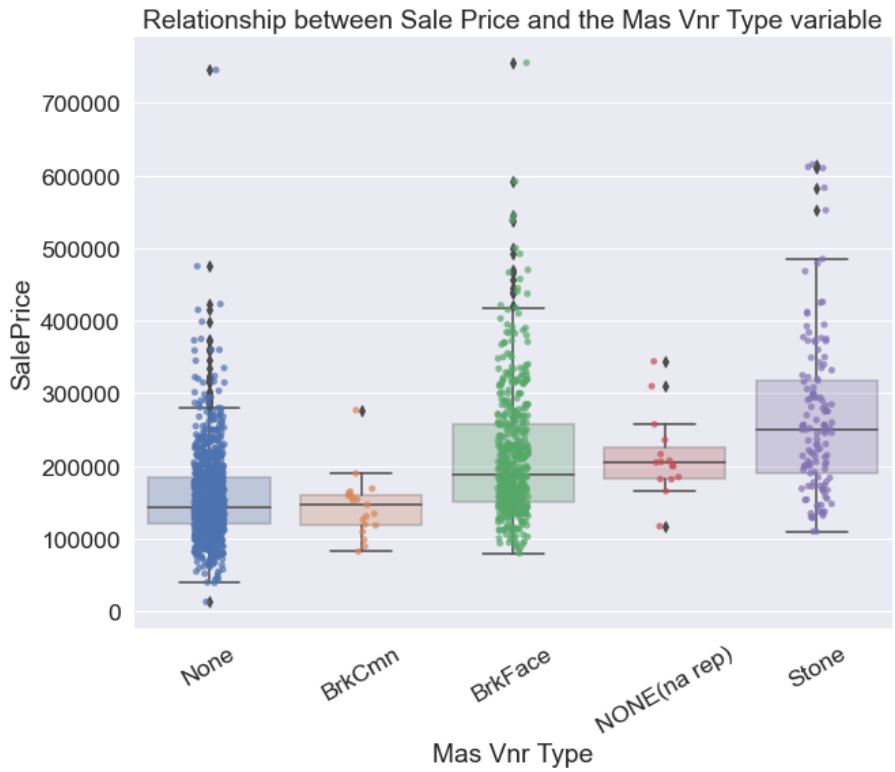

Relationship between Sale Price and the Overall Qual variable
SalePrice
0
100000
200000
300000
400000
500000
600000
700000
1
2
3
4
5
6
7
8
9
10
Overall Qual

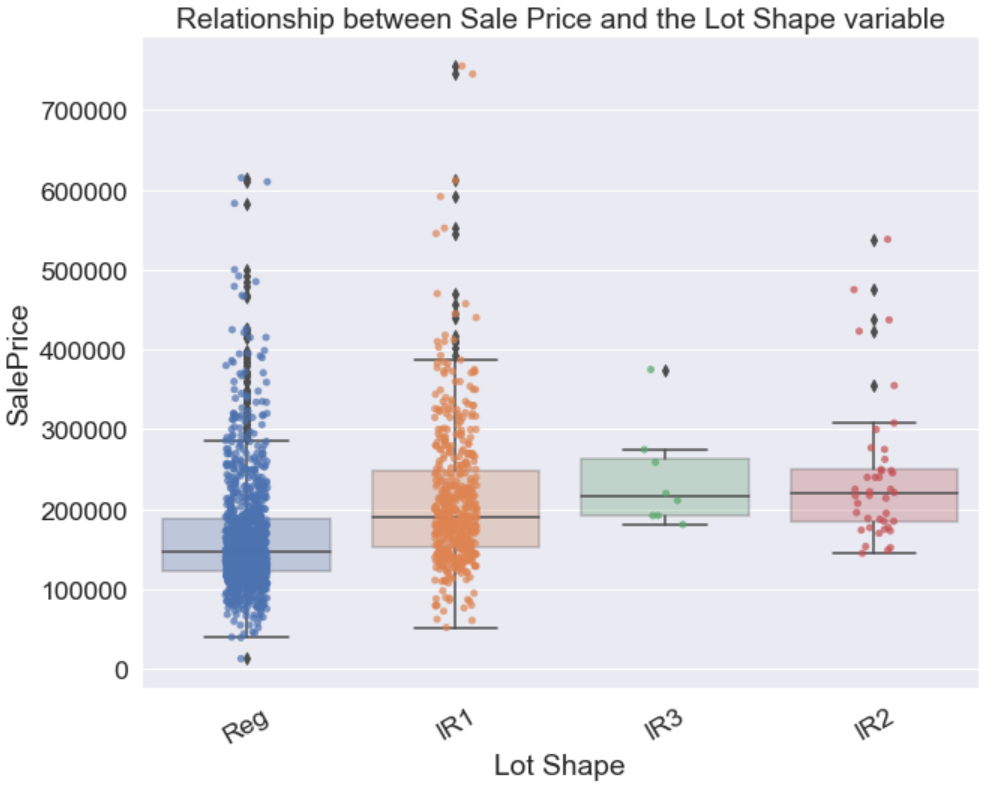

Relationship between Sale Price and the Overall Cond variable
SalePrice
0
100000
200000
300000
400000
500000
600000
700000
1
2
3
4
7
6
8
5
9
Overall Cond

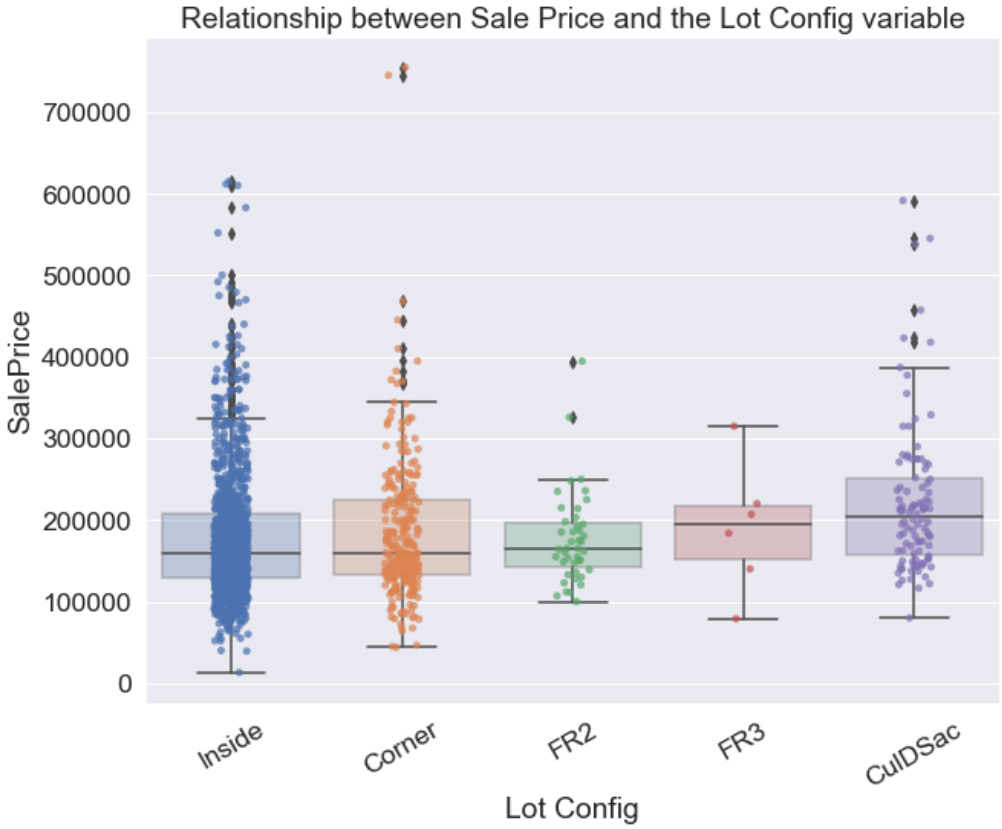

Relationship between Sale Price and the MS Zoning variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
A (agr)
I (all)
C (all)
RM
RH
RL
FV
MS Zoning


Relationship between Sale Price and the Mo Sold variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
6
8
5
4
3
12
10
11
2
7
9
1
Mo Sold

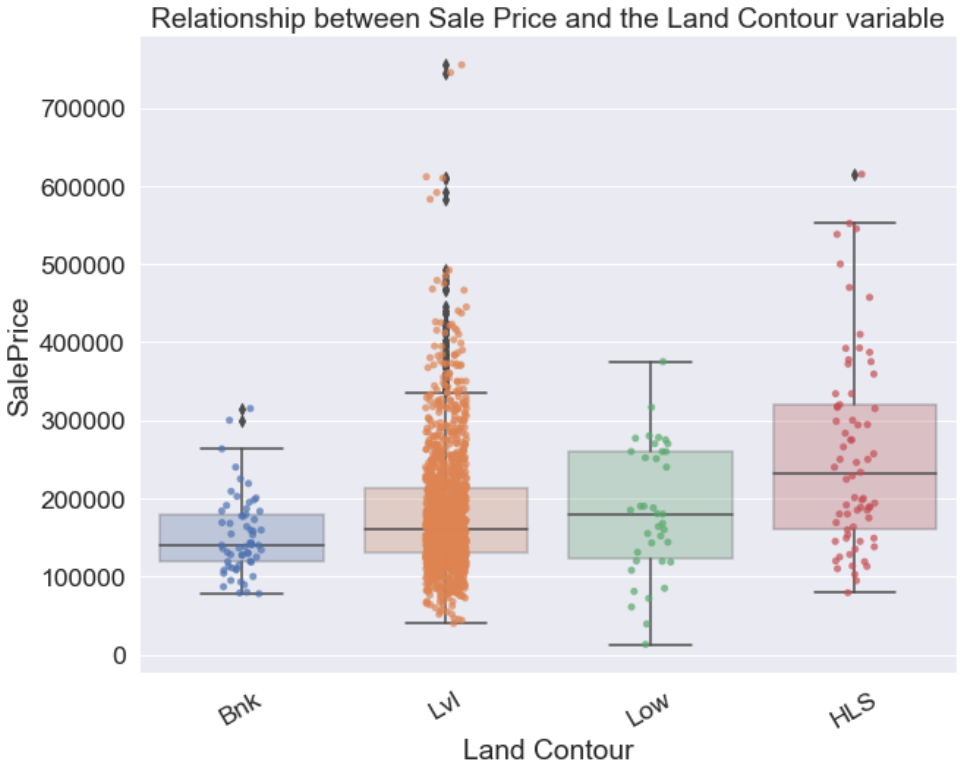

Relationship between Sale Price and the Misc Feature variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Shed
Gar2
NONE(na rep)
Elev
Othr
TenC
Misc Feature

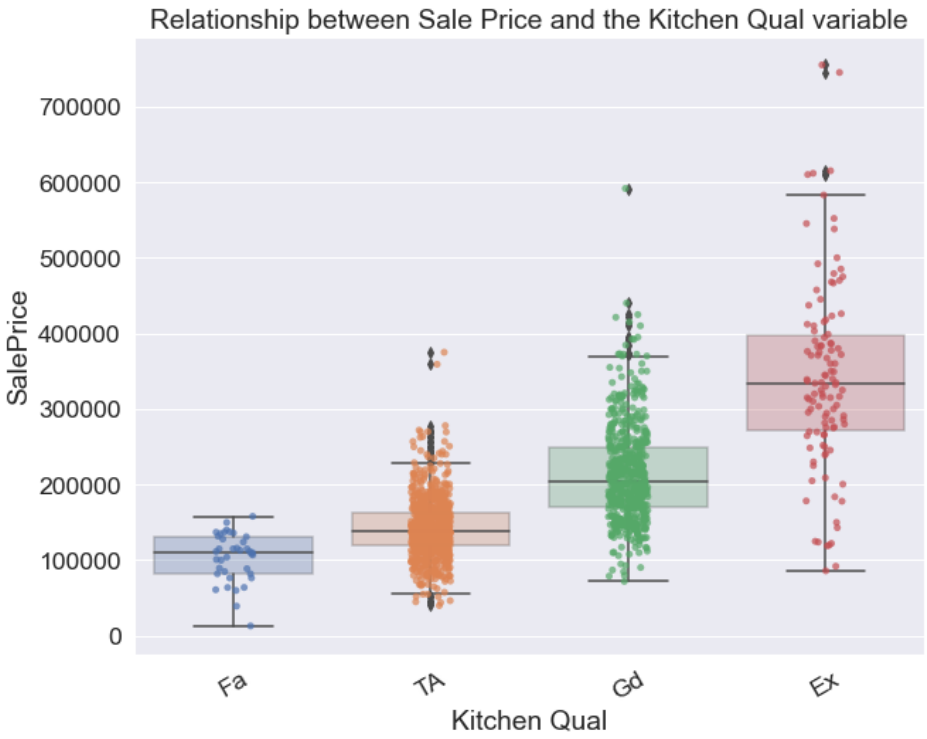

Relationship between Sale Price and the Mas Vnr Type variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
None
BrkCmn
BrkFace
NONE(na rep)
Stone
Mas Vnr Type


Relationship between Sale Price and the Lot Shape variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Reg
IR1
IR3
IR2
Lot Shape

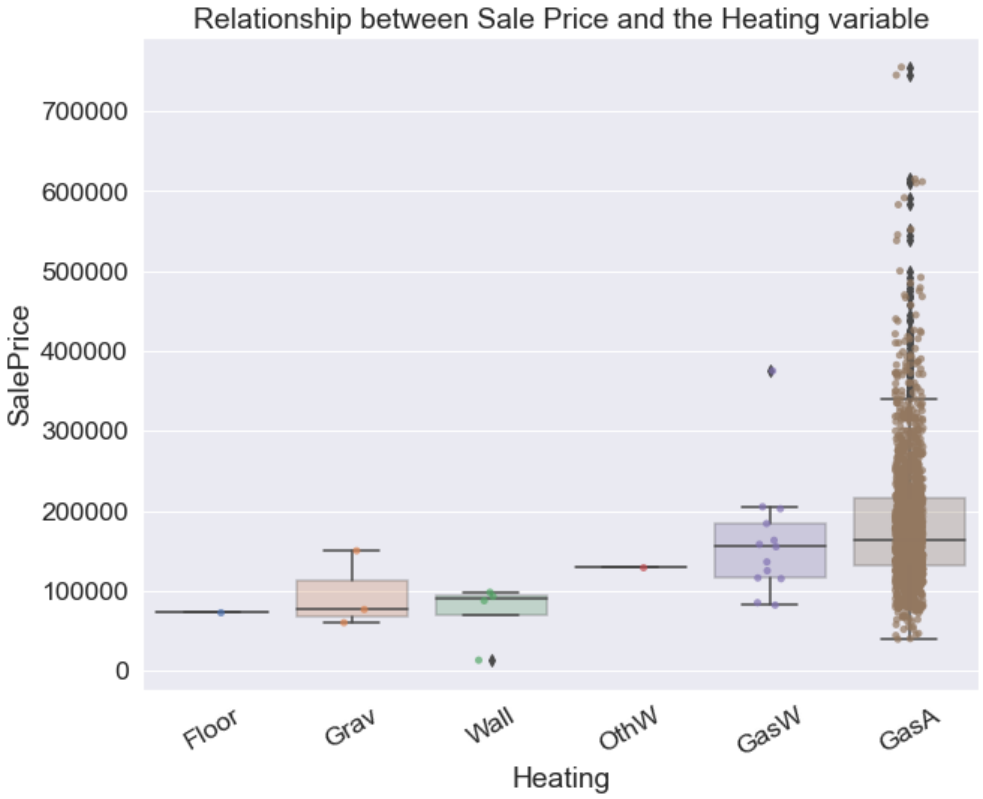

Relationship between Sale Price and the Lot Config variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Inside
Corner
FR2
FR3
CulDSac
Lot Config

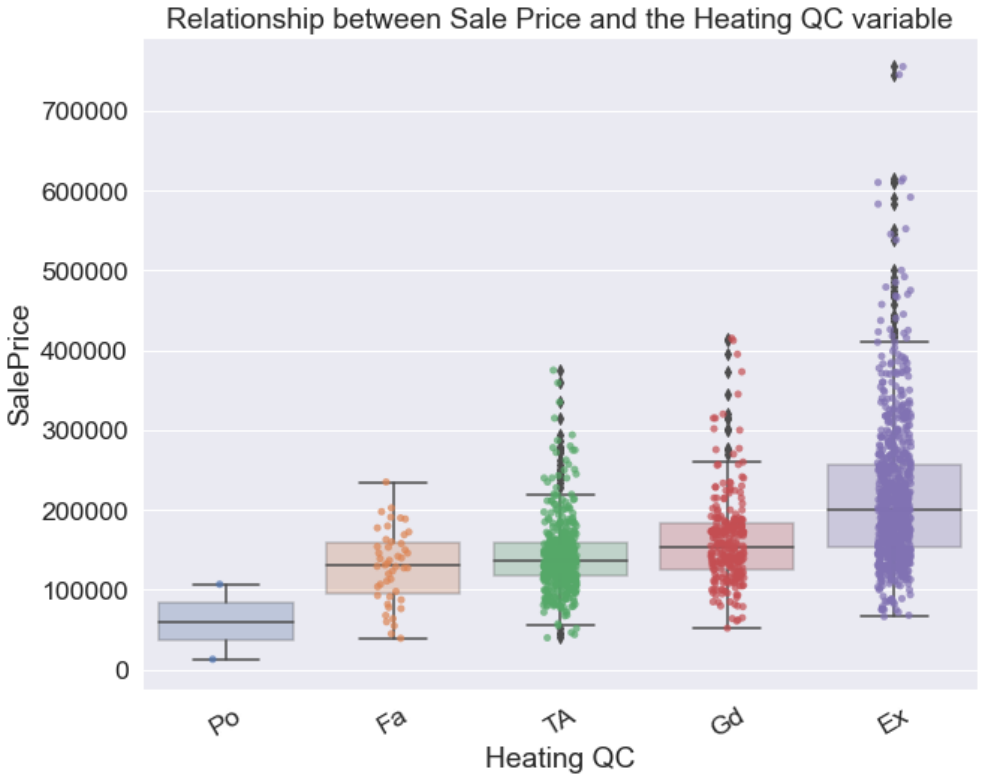

Relationship between Sale Price and the Land Slope variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Gtl
Sev
Mod
Land Slope

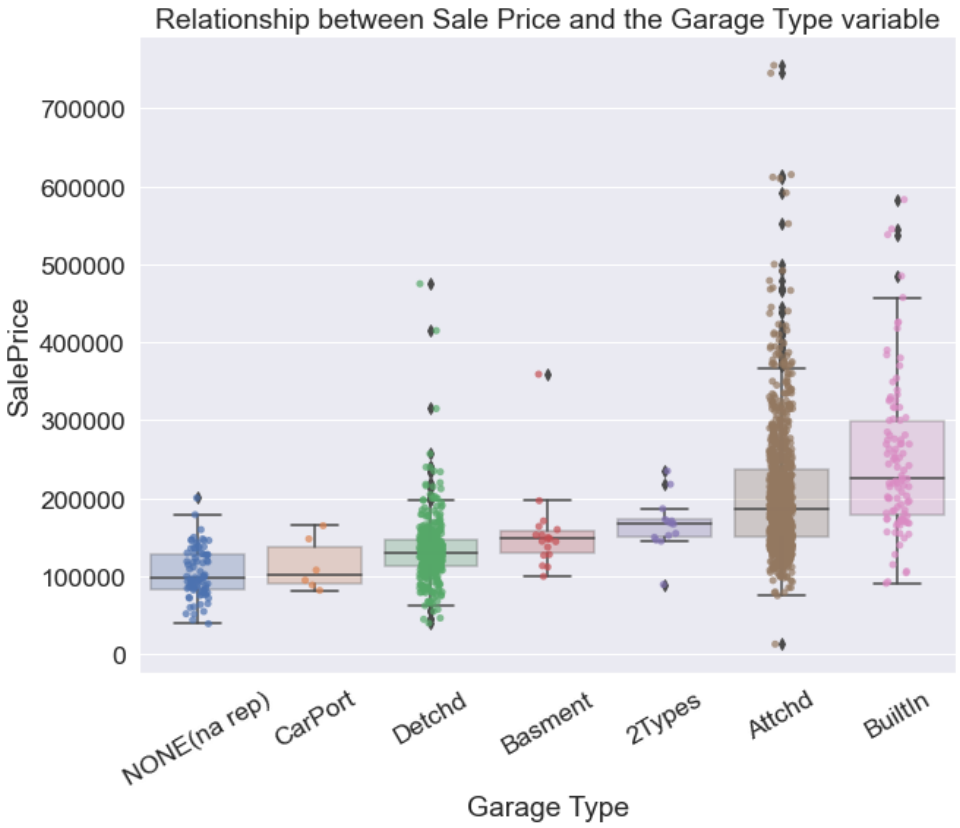

Relationship between Sale Price and the Land Contour variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Bnk
Lvl
Low
HLS
Land Contour

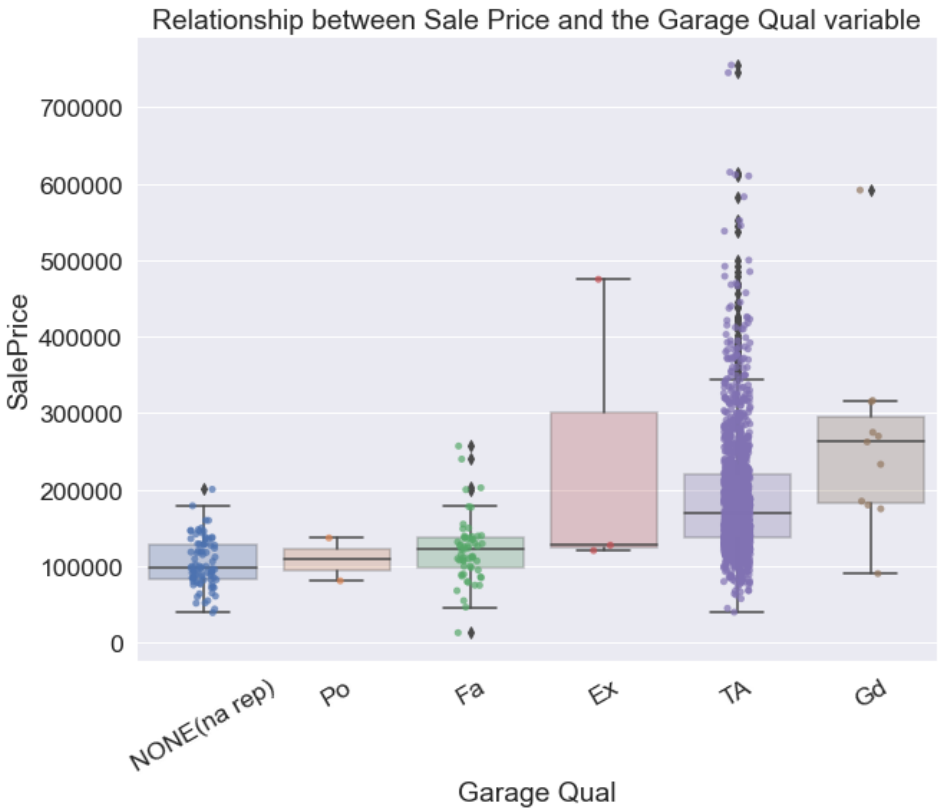

Relationship between Sale Price and the Kitchen Qual variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Fa
TA
Gd
Ex
Kitchen Qual

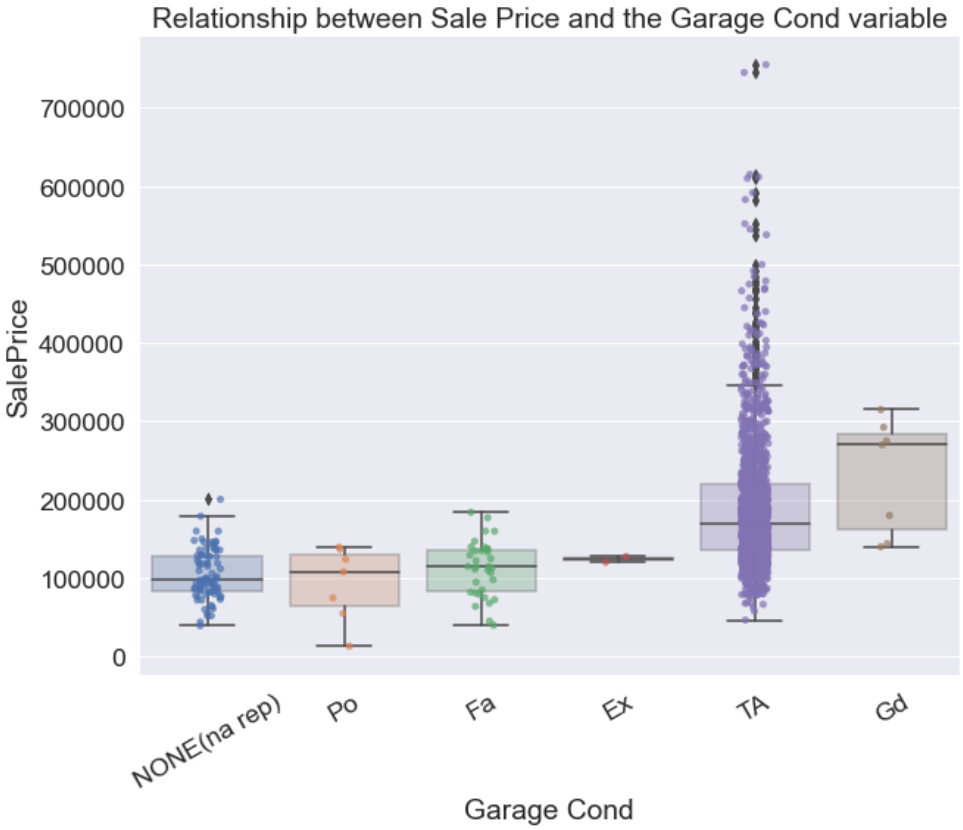

Relationship between Sale Price and the House Style variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
1.5Unf
1.5Fin
SFoyer
1Story
2.5Unf
SLvl
2Story
2.5Fin
House Style

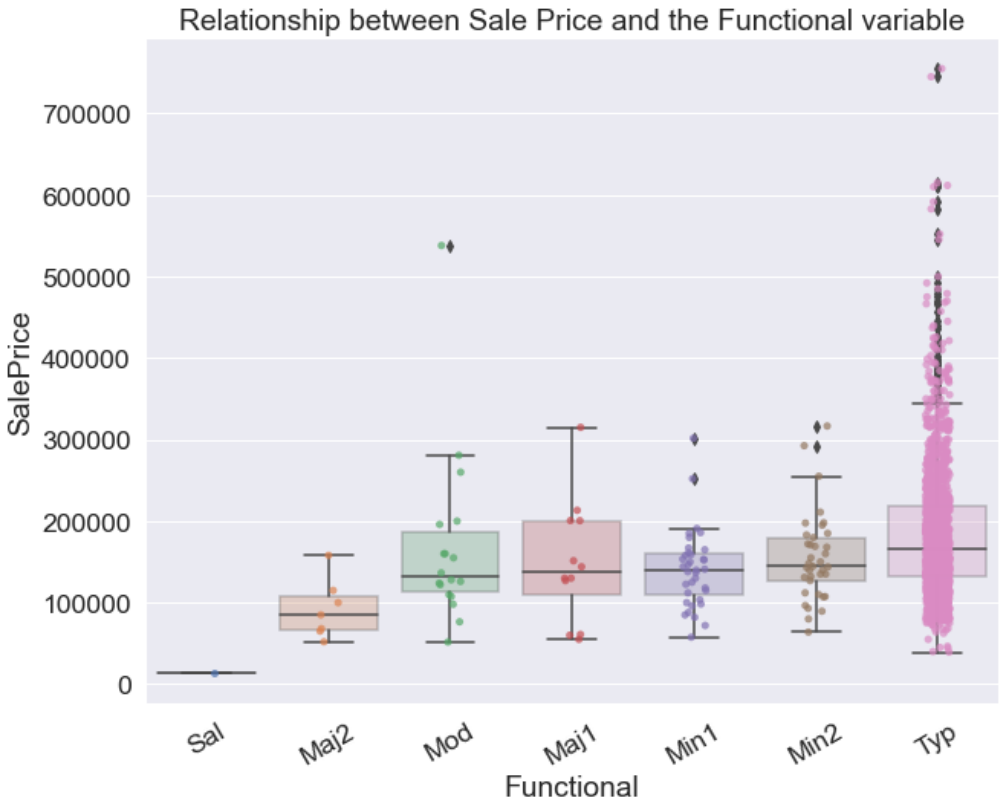

Relationship between Sale Price and the Heating variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Floor
Grav
Wall
OthW
GasW
GasA
Heating

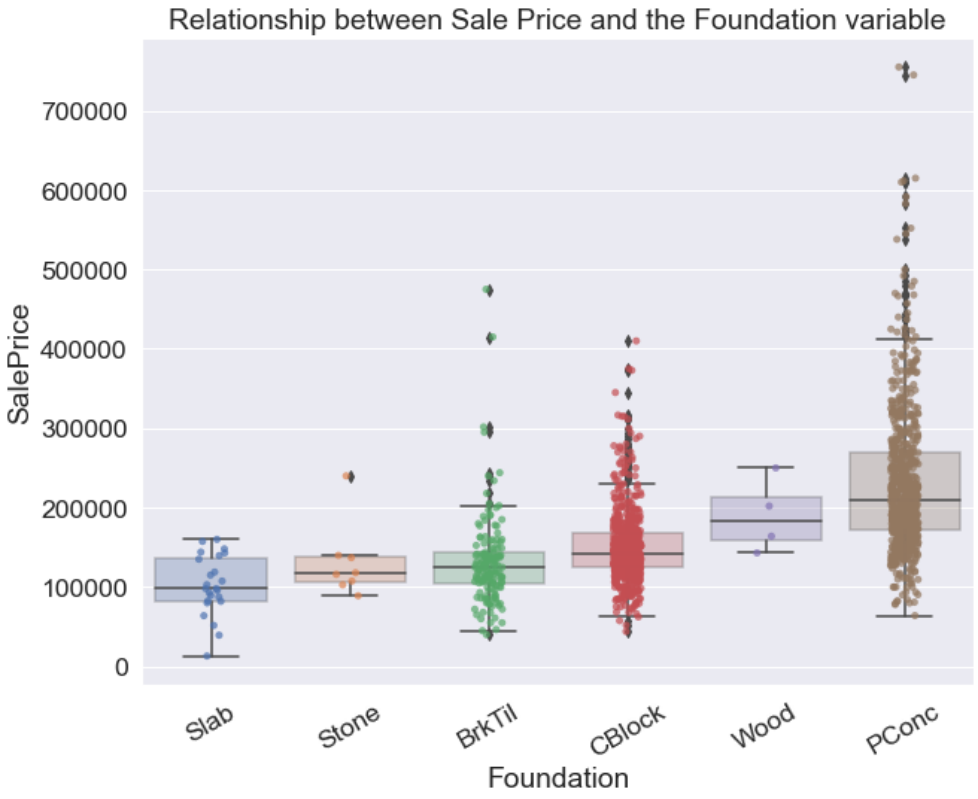

Relationship between Sale Price and the Heating QC variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Po
Fa
TA
Gd
Ex
Heating QC

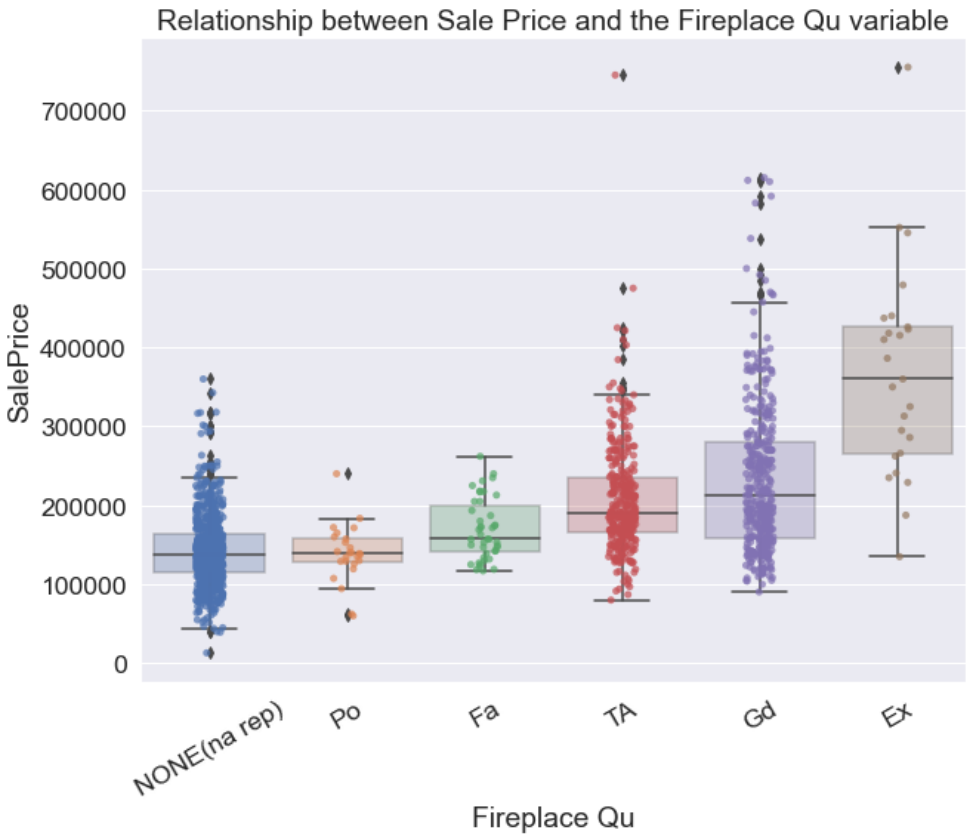

Relationship between Sale Price and the Garage Type variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
NONE(na rep)
CarPort
Detchd
Basment
2Types
Attchd
BuiltIn
Garage Type

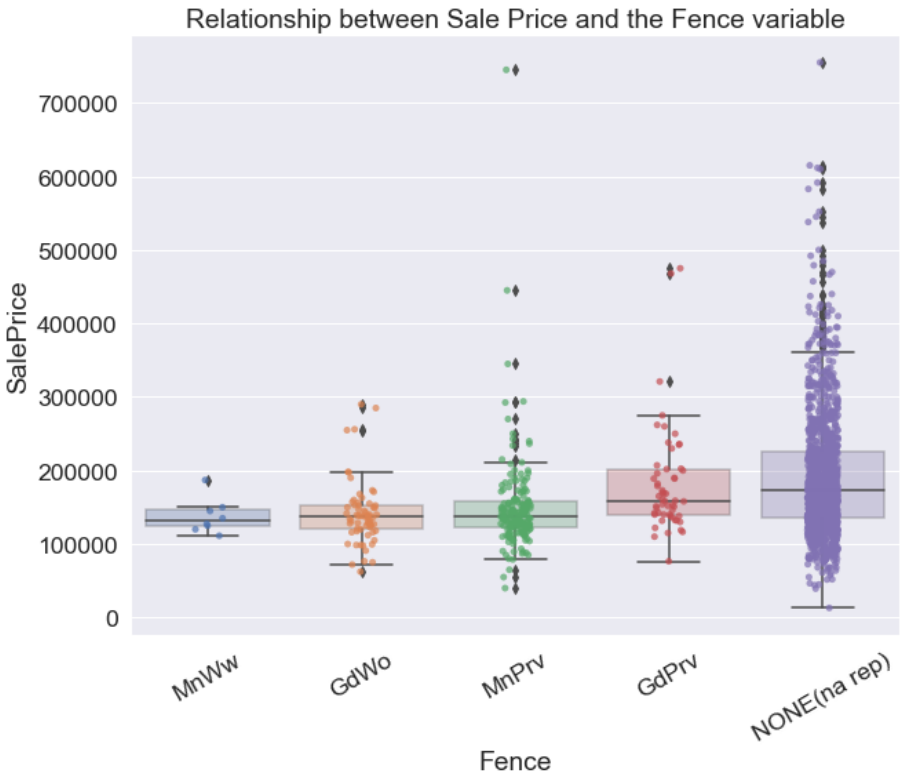

Relationship between Sale Price and the Garage Qual variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
NONE(na rep)
Po
Fa
Ex
TA
Gd
Garage Qual


Relationship between Sale Price and the Garage Cond variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
NONE(na rep)
Po
Fa
Ex
TA
Gd
Garage Cond

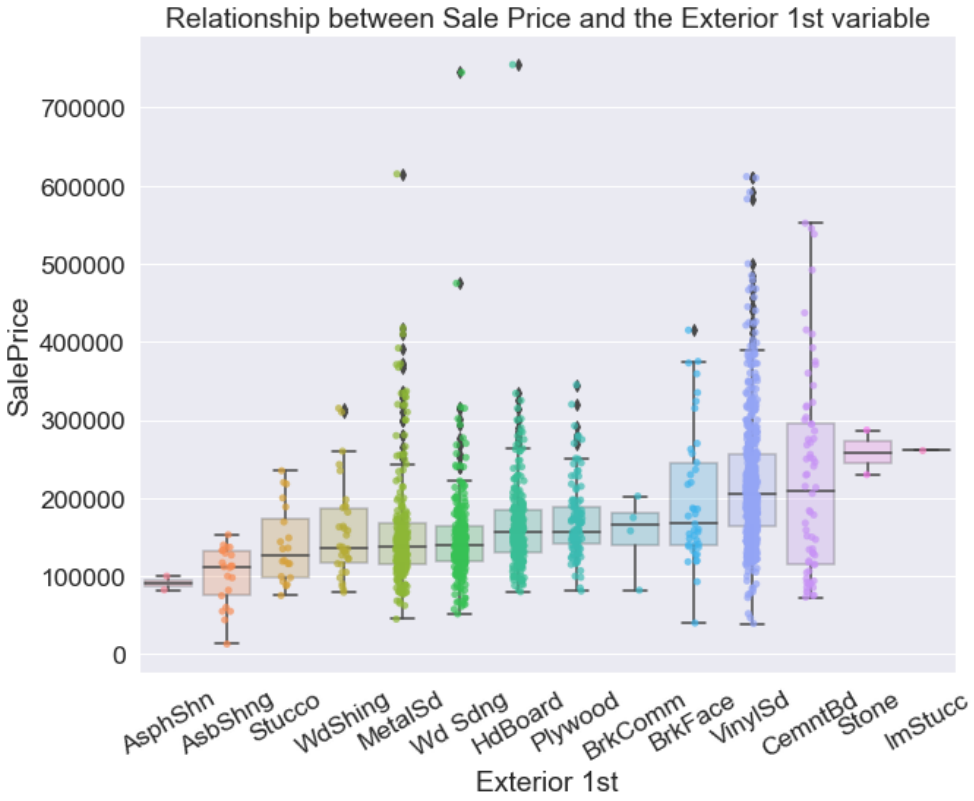

Relationship between Sale Price and the Functional variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Sal
Maj2
Mod
Maj1
Min1
Min2
Typ
Functional

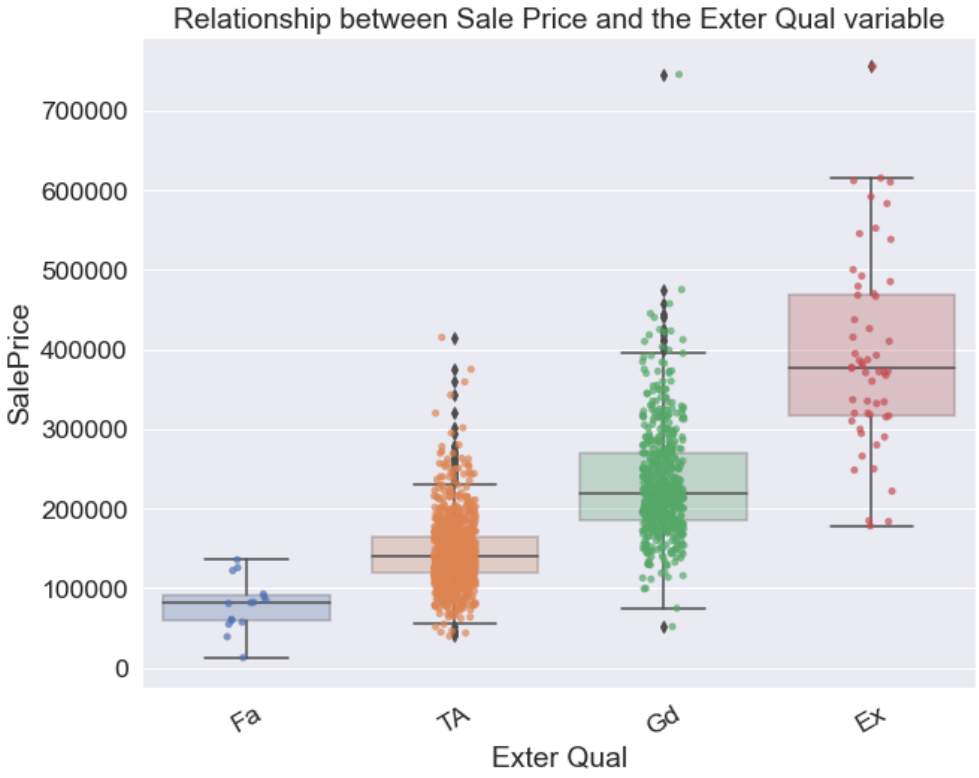

Relationship between Sale Price and the Foundation variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Slab
Stone
BrkTil
CBlock
Wood
PConc
Foundation


Relationship between Sale Price and the Fireplace Qu variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
NONE(na rep)
Po
Fa
TA
Gd
Ex
Fireplace Qu

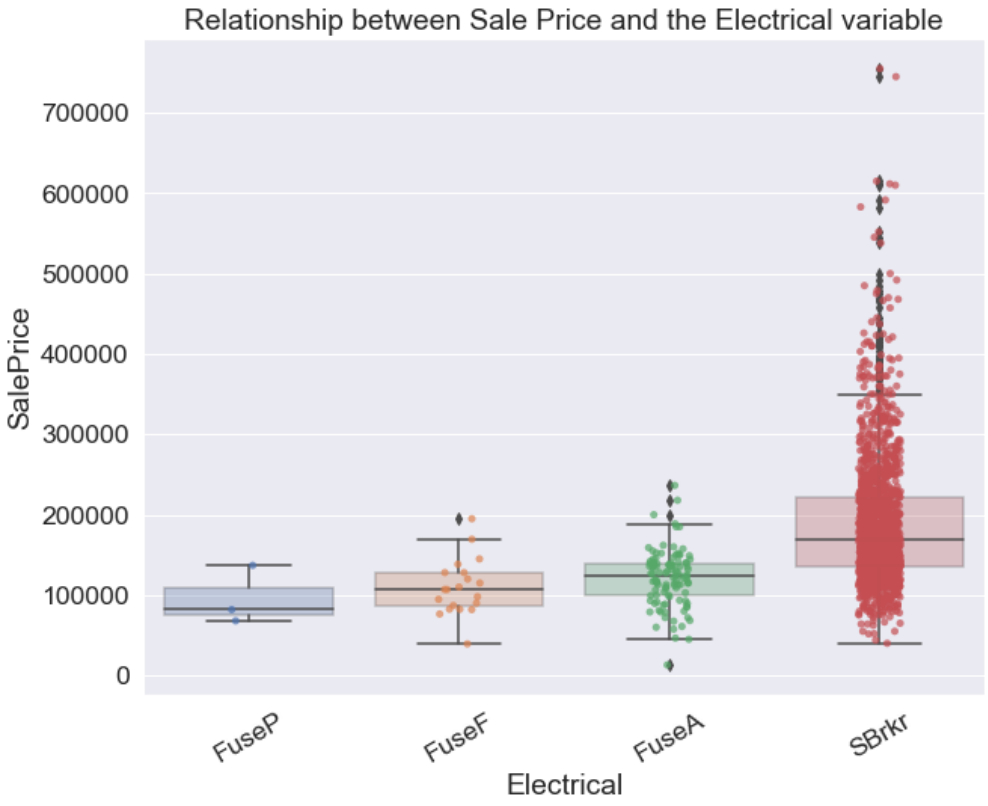

Relationship between Sale Price and the Fence variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
MnWw
GdWo
MnPrv
GdPrv
NONE(na rep)
Fence

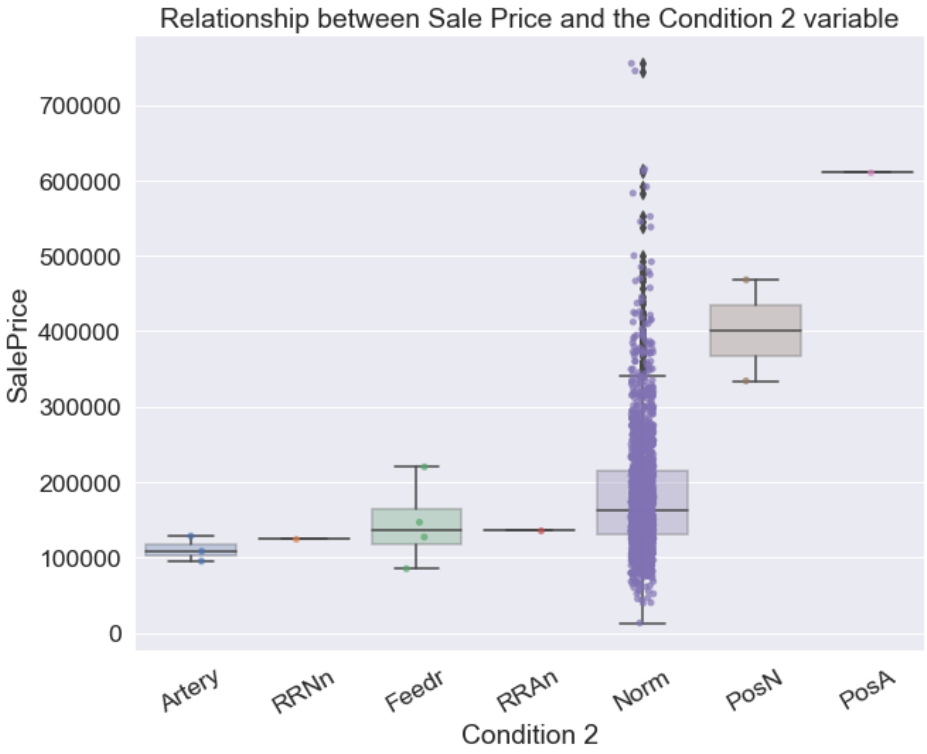

Relationship between Sale Price and the Exterior 2nd variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
CBlock
AsphShn
AsbShng
Stucco
Wd Sdng
MetalSd
Brk Cmn
Wd Shng
Plywood
HdBoard
BrkFace
Stone
VinylSd
CmentBd
ImStucc
Other
Exterior 2nd


Relationship between Sale Price and the Exterior 1st variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
AsphShn
AsbShng
Stucco
WdShing
MetalSd
Wd Sdng
HdBoard
Plywood
BrkComm
BrkFace
VinylSd
CemntBd
Stone
ImStucc
Exterior 1st

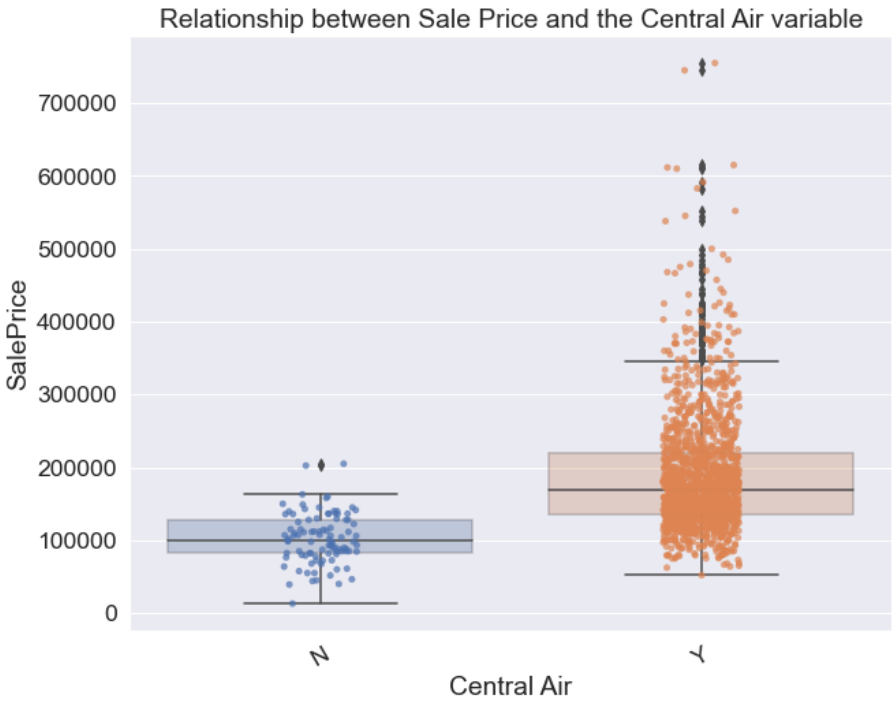

Relationship between Sale Price and the Exter Qual variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Fa
TA
Gd
Ex
Exter Qual

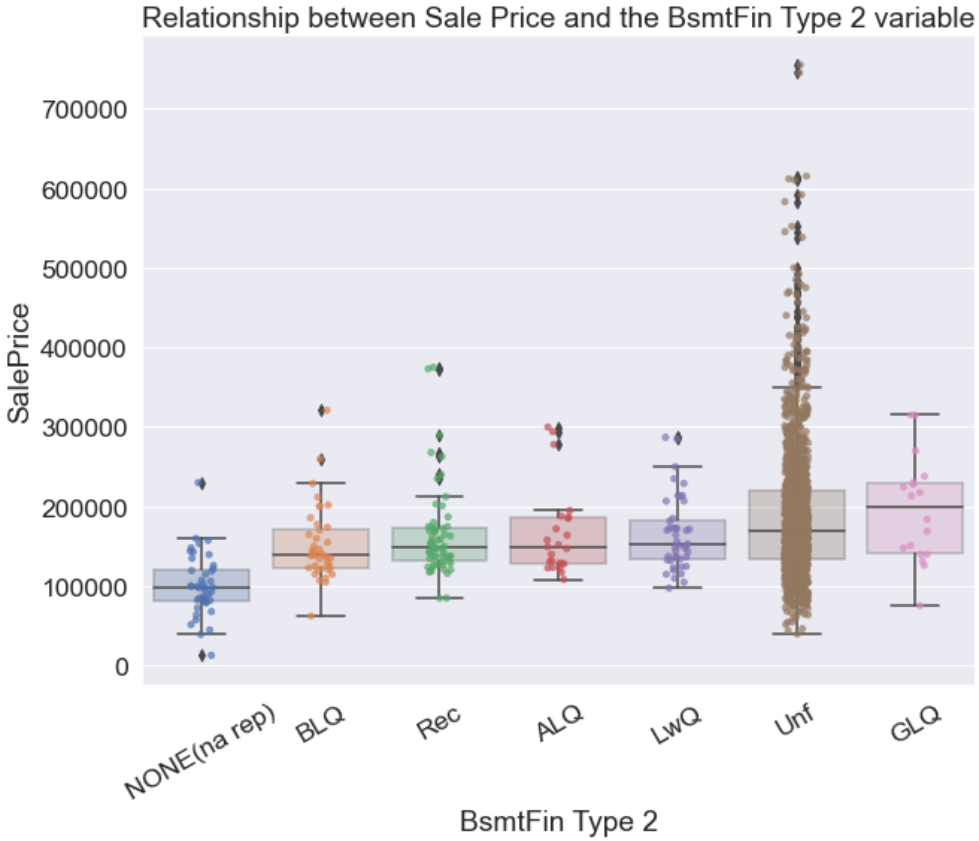

Relationship between Sale Price and the Exter Cond variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Po
Fa
Gd
TA
Ex
Exter Cond

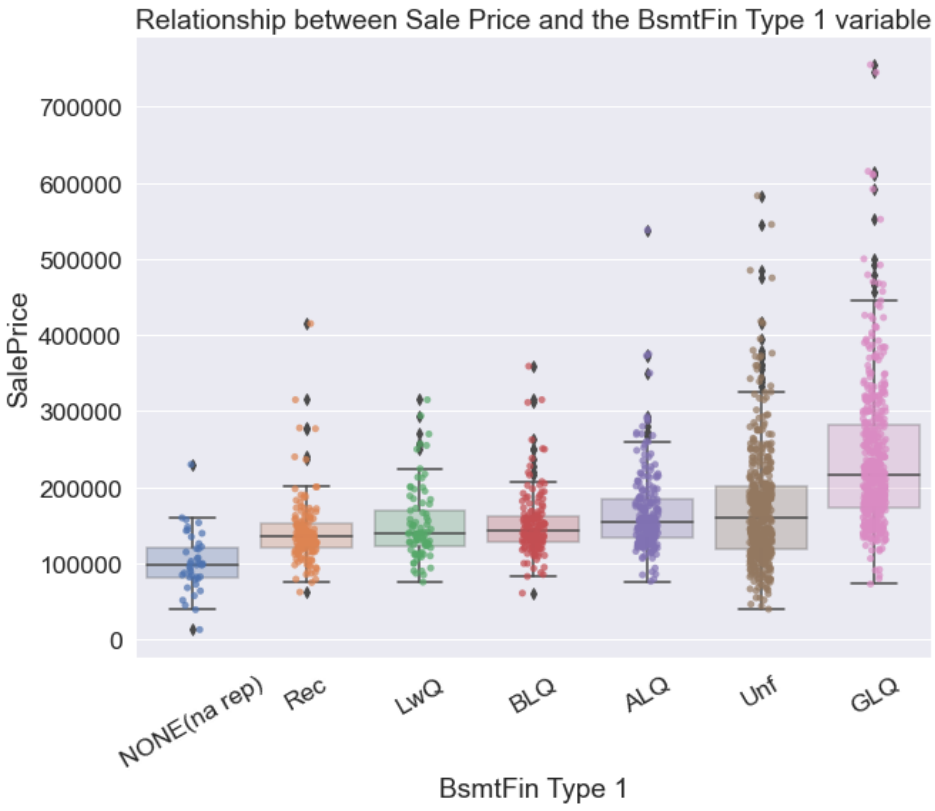

Relationship between Sale Price and the Electrical variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
FuseP
FuseF
FuseA
SBrkr
Electrical

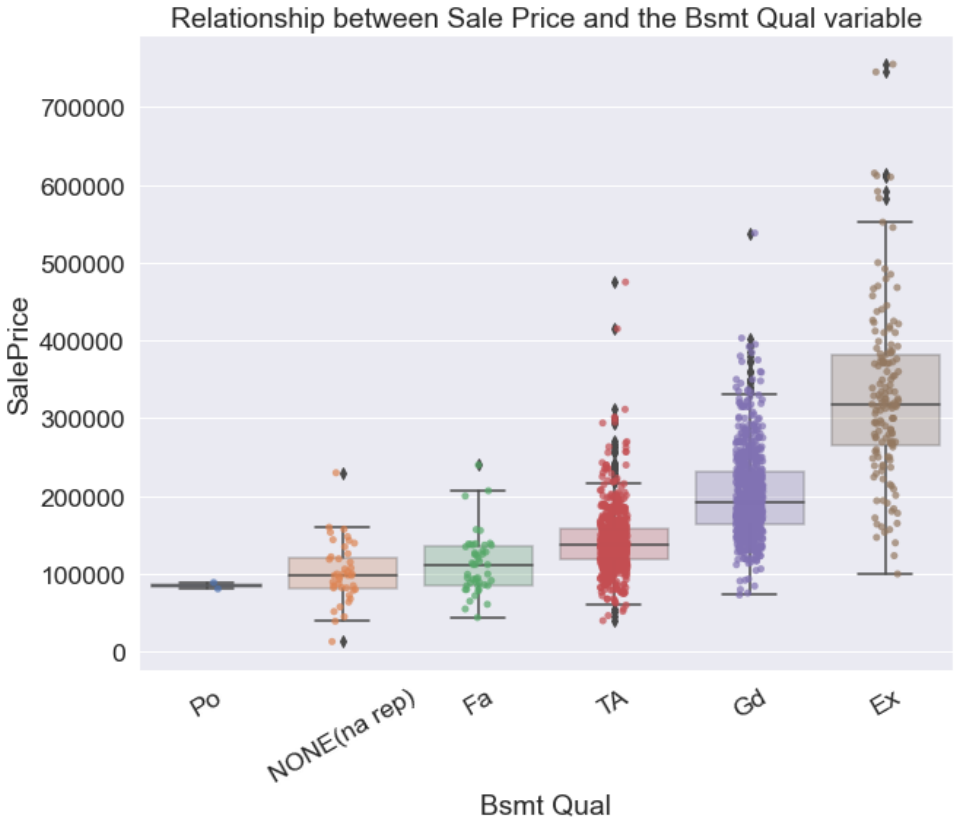

Relationship between Sale Price and the Condition 2 variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Artery
RRNn
Feedr
RRAn
Norm
PosN
PosA
Condition 2

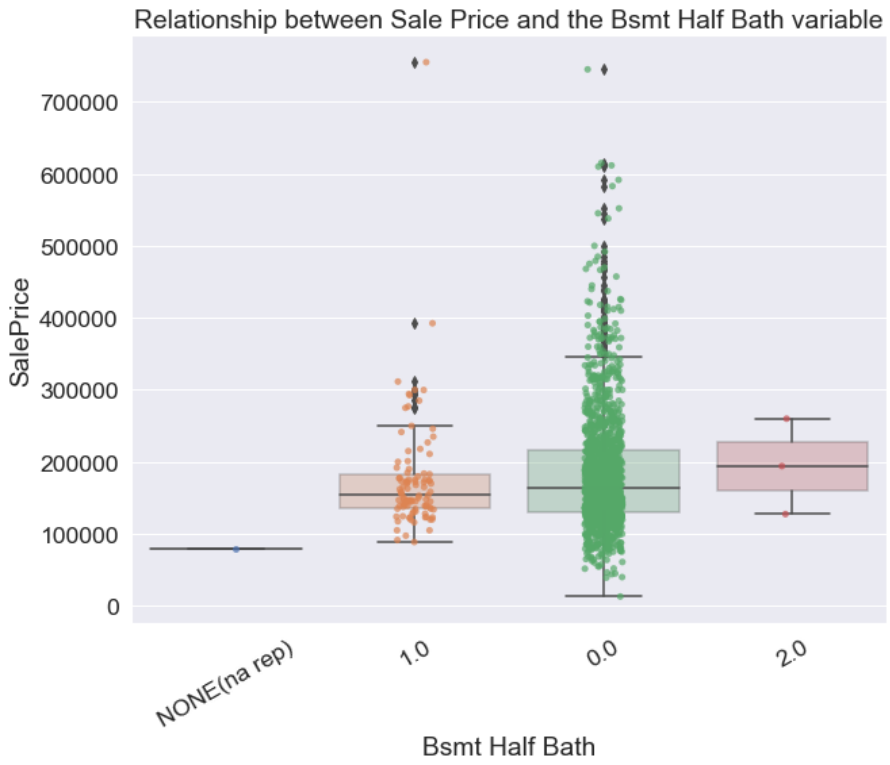

Relationship between Sale Price and the Condition 1 variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Artery
Feedr
RRAe
Norm
RRNe
RRAn
PosA
PosN
RRNn
Condition 1

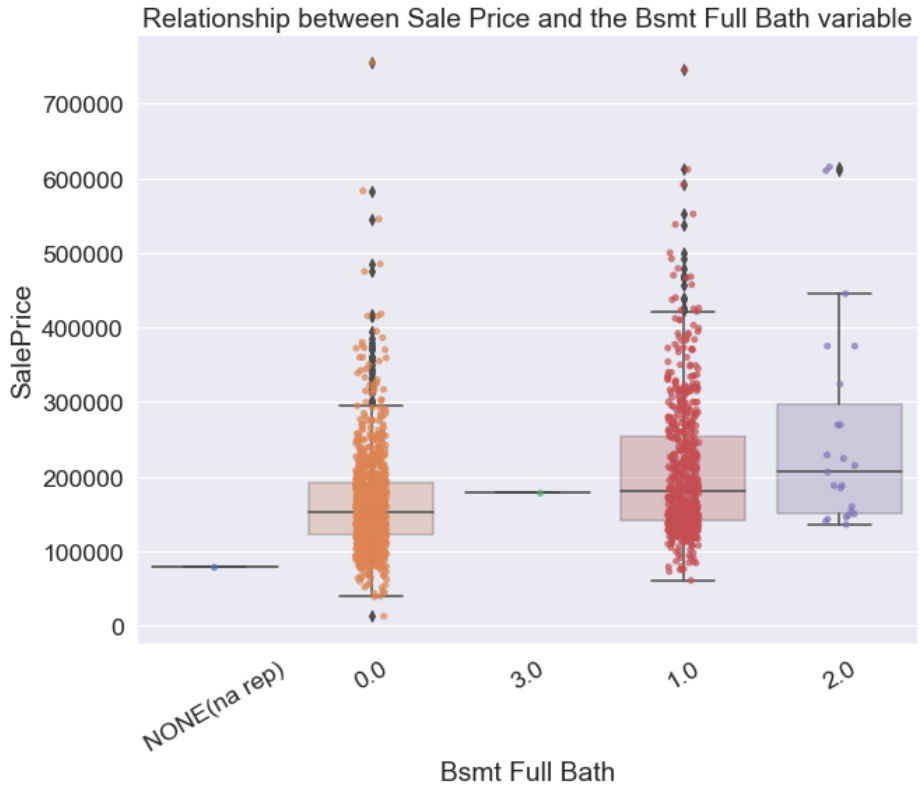

Relationship between Sale Price and the Central Air variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
N
Y
Central Air

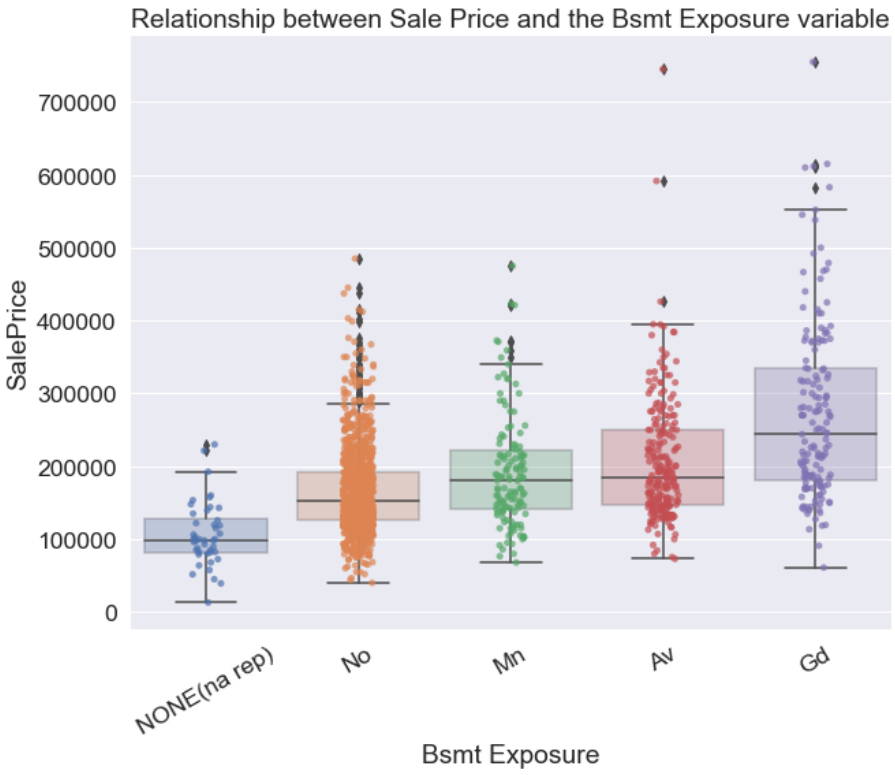

Relationship between Sale Price and the BsmtFin Type 2 variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
NONE(na rep)
BLQ
Rec
ALQ
LwQ
Unf
GLQ
BsmtFin Type 2

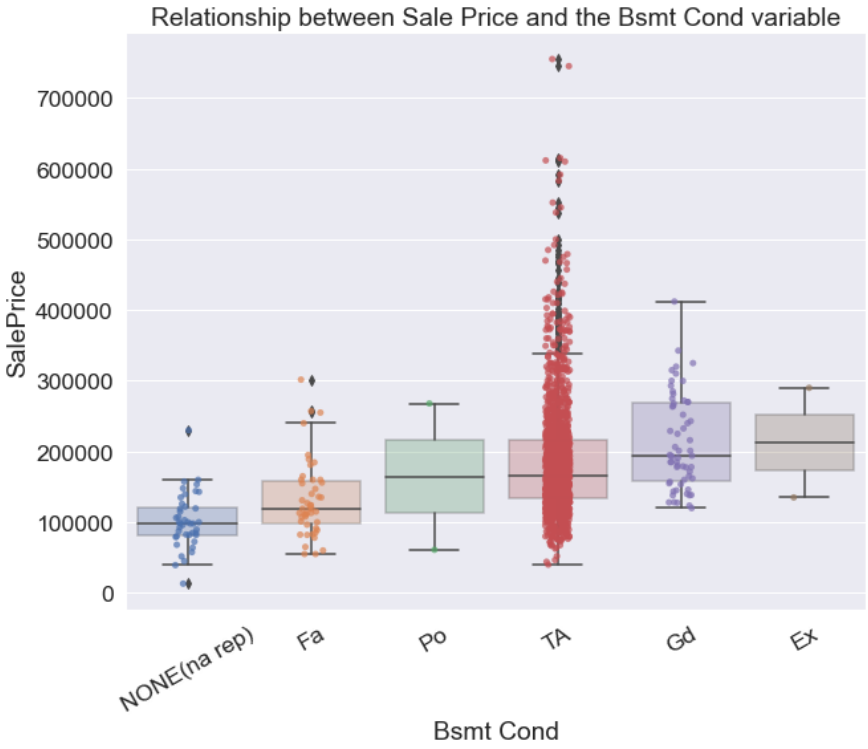

Relationship between Sale Price and the BsmtFin Type 1 variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
NONE(na rep)
Rec
LwQ
BLQ
ALQ
Unf
GLQ
BsmtFin Type 1

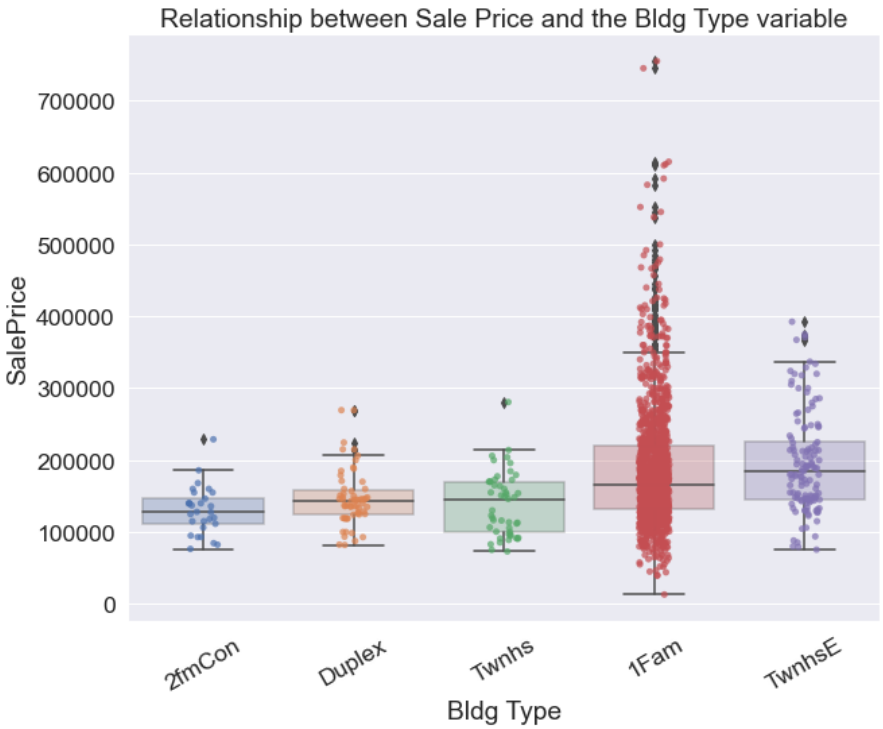

Relationship between Sale Price and the Bsmt Qual variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Po
NONE(na rep)
Fa
TA
Gd
Ex
Bsmt Qual

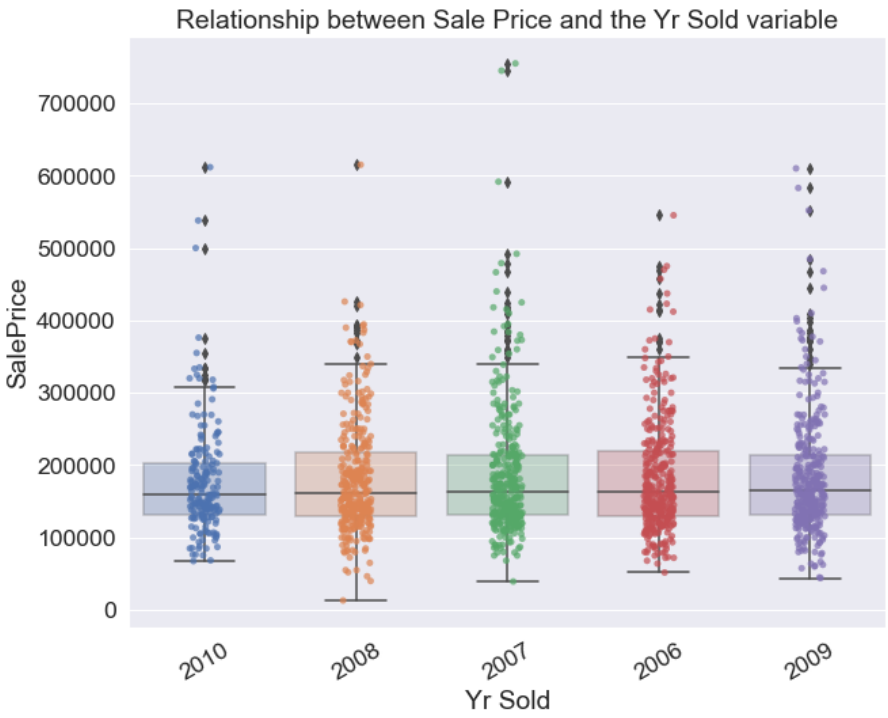

Relationship between Sale Price and the Bsmt Half Bath variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
NONE(na rep)
1.0
0.0
2.0
Bsmt Half Bath


Relationship between Sale Price and the Bsmt Full Bath variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
NONE(na rep)
0.0
3.0
1.0
2.0
Bsmt Full Bath


Relationship between Sale Price and the Bsmt Exposure variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
NONE(na rep)
No
Mn
Av
Gd
Bsmt Exposure


Relationship between Sale Price and the Bsmt Cond variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
NONE(na rep)
Fa
Po
TA
Gd
Ex
Bsmt Cond


Relationship between Sale Price and the Bldg Type variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
2fmCon
Duplex
Twnhs
1Fam
TwnhsE
Bldg Type


Relationship between Sale Price and the Yr Sold variable
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
2010
2008
2007
2006
2009
Yr Sold

Appendix v

```
count       312.0000
mean     419407.3718
std      132443.0593
min      161400.0000
25%      332625.0000
50%      412950.0000
75%      488775.0000
max      971800.0000
Name: Number of Visitors, dtype: float64
```