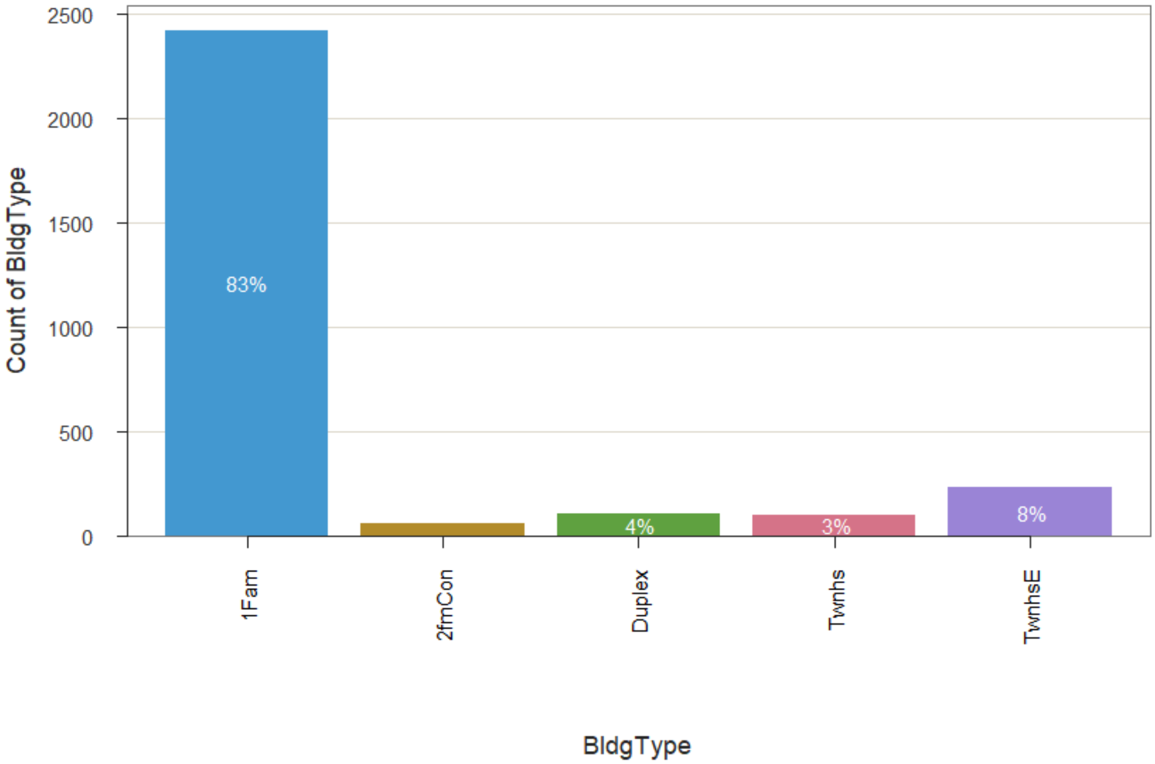
Reed Ballesteros
MSDS-410-DL, Summer 2022
Dr. Mickelson
6/26/2022

Modeling Assignment #1: Getting to Know Your Data - Exploratory Data Analysis (EDA)

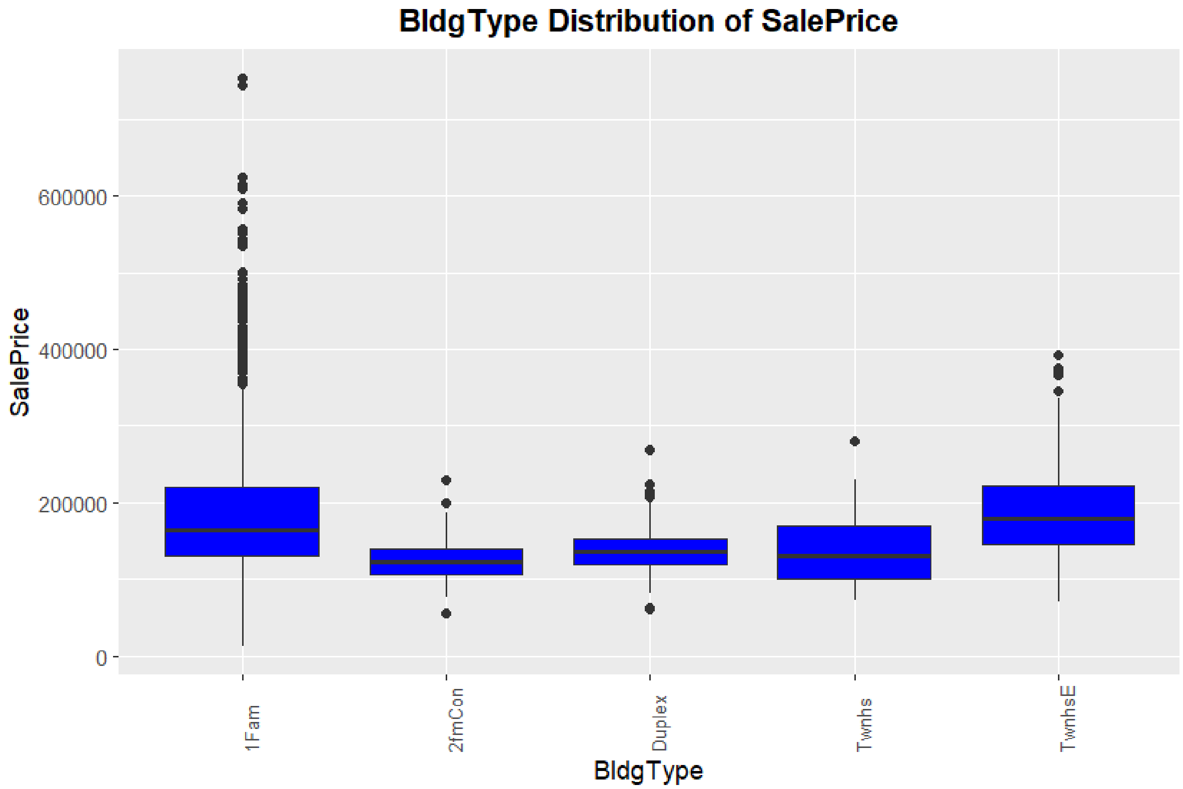
Given the Ames Iowa dataset, we will attempt to create regression models to predict a value of a home from the initial 82 variables available. The variables represent numerous properties of a typical home in the area, such as lot, floor, and garage & basement area; bedroom, kitchen, and bathroom details; neighborhood and amenities (heating, deck, pool, etc.) information; and overall quality ratings. The dataset should provide the insight to create regression models for predicting a home value.

1. Sample Definition

From the bar chart below, we can see that single family homes (noted as BldgType = ‘1Fam’) make up an overwhelming 83% (2425 of the 2930 observations) of home types listed in the Ames dataset. Because of this, we will define the drop condition of our sample population of interest to only include single family homes in our dataset.



From the box plot below, we also see that while single family homes do yield a higher sale price, it also represents a wider range of home values as well.



2. Data Quality Check

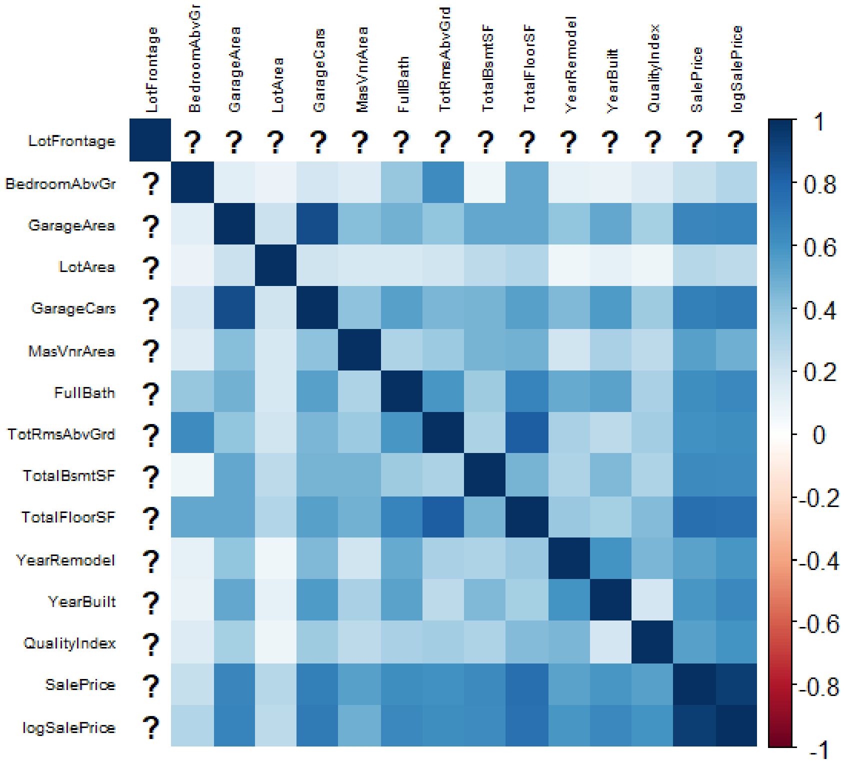
Though the Ames dataset has 82 variables, we will pick twenty that have a relative correlation to SalePrice. Quantitative variables for the top twenty include the following (definitions are from the Ames Housing data documentation):

- LotFrontage - Linear feet of street connected to property
- LotArea - Lot size in square feet
- GarageArea - Size of garage in square feet
- GarageCars - Size of garage in car capacity
- MasVnrArea - Masonry veneer area in square feet
- FullBath - Full bathrooms above grade
- TotRmsAbvGrd - Total rooms above grade (does not include bathrooms)
- BedroomAbvGr - Bedrooms above grade (does NOT include basement bedrooms)
- TotalBsmtSF - Total square feet of basement area
- YearRemodel - Remodel date (same as construction date if no remodeling or additions)
- YearBuilt - Original construction date

Two new variables are based on transformations of other variables to create a greater correlation with SalePrice:

- TotalFloorSF = FirstFloorSF + SecondFlrSF – combination of FirstFloorSF (First Floor square feet) and SecondFlrSF (Second floor square feet, if available)
- QualityIndex = OverallQual * OverallCond – index based on OverallQual (Rates the overall material and finish of the house) and OverallCond (Rates the overall condition of the house)

A correlation table of the quantitative variables above along with SalePrice and logSalePrice is provided below:

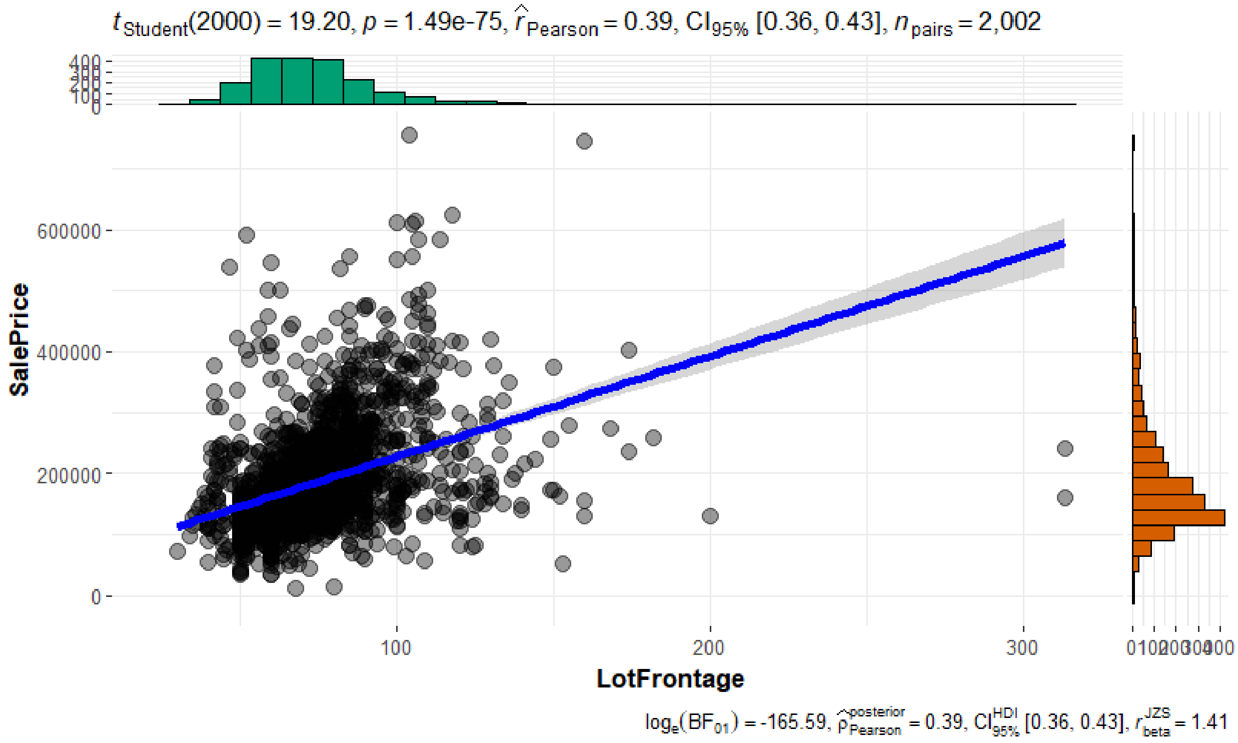


We can see that BedroomAbvGr , FullBath, YearRemodel, YearBuilt, and QualityIndex have a slightly larger correlation with log(SalePrice) compared to SalePrice.

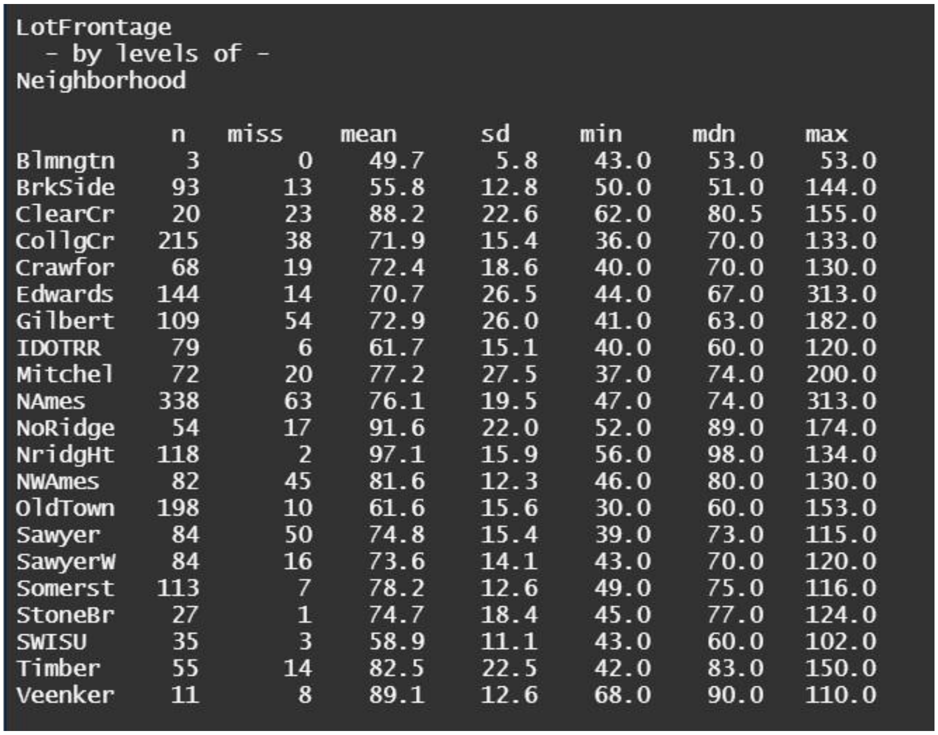
Data cleanup needed to be performed on the following variables due to some rows having null values:

- MasVnrArea
- GarageArea
- GarageCars
- LotFrontage

While homes with null MasVnrArea, GarageArea, GarageCars values just needed to fill in 0 to denote they do not have mason veneer or garages, a null LotFrontage cannot simply be 0 by default. The correlation table above shows ‘?’ in respect of correlation to the other quantitative variables due to null LotFontage values. We should not drop homes with null LotFrontage because doing so will remove almost 20% of the single family home sample. We should not ignore LotFrontage as well because for homes that do have it shows a relative correlation to SalePrice (0.39):



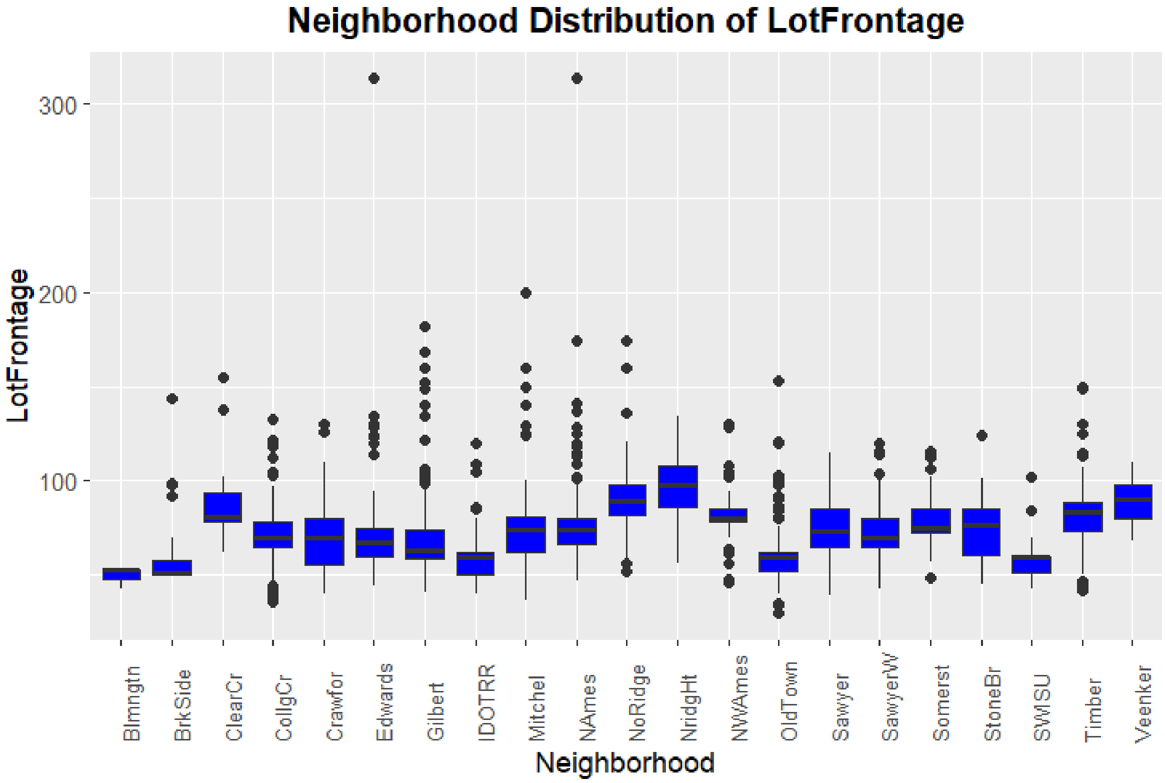
To remedy the issue, we will fill homes with null LotFrontage with the mean value based on the neighborhood (estimated to the nearest integer).



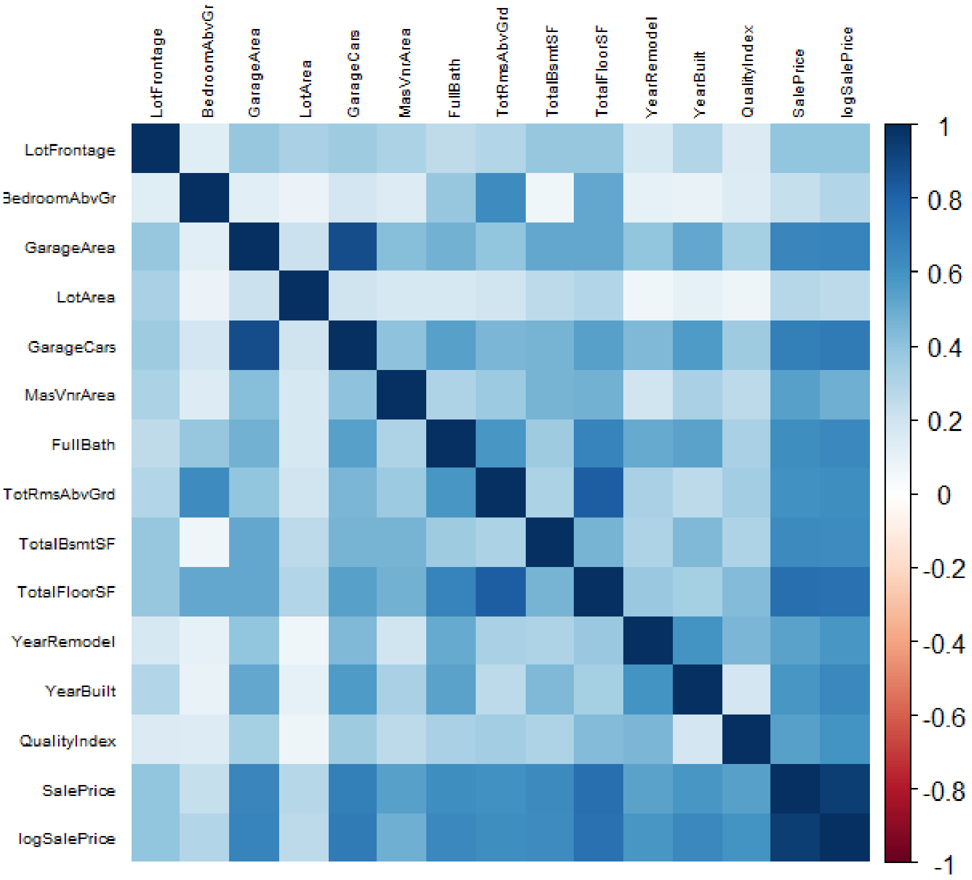
```
LotFrontage
  - by levels of -
Neighborhood

           n   miss    mean     sd     min     mdn     max
Blmngtn    3      0    49.7    5.8    43.0    53.0    53.0
BrkSide   93     13    55.8   12.8    50.0    51.0   144.0
ClearCr   20     23    88.2   22.6    62.0    80.5   155.0
CollgCr  215     38    71.9   15.4    36.0    70.0   133.0
Crawfor   68     19    72.4   18.6    40.0    70.0   130.0
Edwards  144     14    70.7   26.5    44.0    67.0   313.0
Gilbert  109     54    72.9   26.0    41.0    63.0   182.0
IDOTRR    79      6    61.7   15.1    40.0    60.0   120.0
Mitchel   72     20    77.2   27.5    37.0    74.0   200.0
NAmes    338     63    76.1   19.5    47.0    74.0   313.0
NoRidge   54     17    91.6   22.0    52.0    89.0   174.0
NridgHt  118      2    97.1   15.9    56.0    98.0   134.0
NWAmes    82     45    81.6   12.3    46.0    80.0   130.0
OldTown  198     10    61.6   15.6    30.0    60.0   153.0
Sawyer    84     50    74.8   15.4    39.0    73.0   115.0
SawyerW   84     16    73.6   14.1    43.0    70.0   120.0
Somerst  113      7    78.2   12.6    49.0    75.0   116.0
StoneBr   27      1    74.7   18.4    45.0    77.0   124.0
SWISU     35      3    58.9   11.1    43.0    60.0   102.0
Timber    55     14    82.5   22.5    42.0    83.0   150.0
Veenker   11      8    89.1   12.6    68.0    90.0   110.0
```

The distribution of LotFrontage by Neighborhood is shown below.



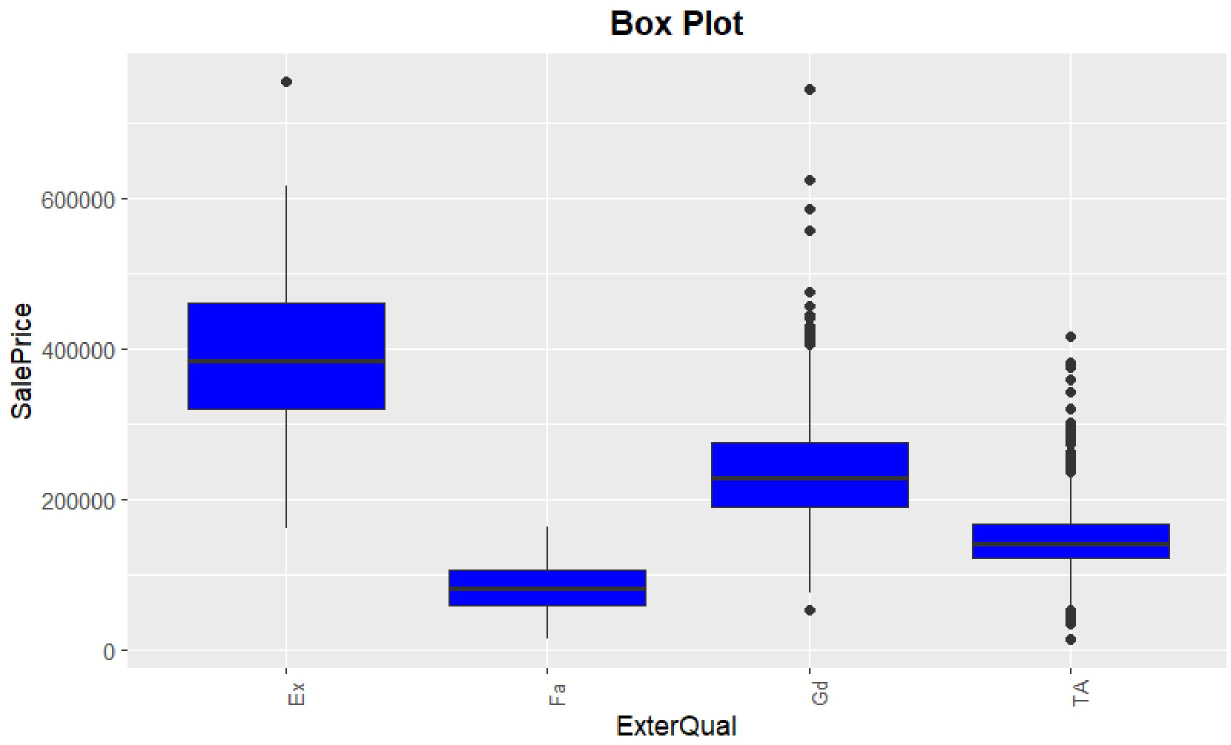
After the LotFrontage cleanup the correlation table is updated below:



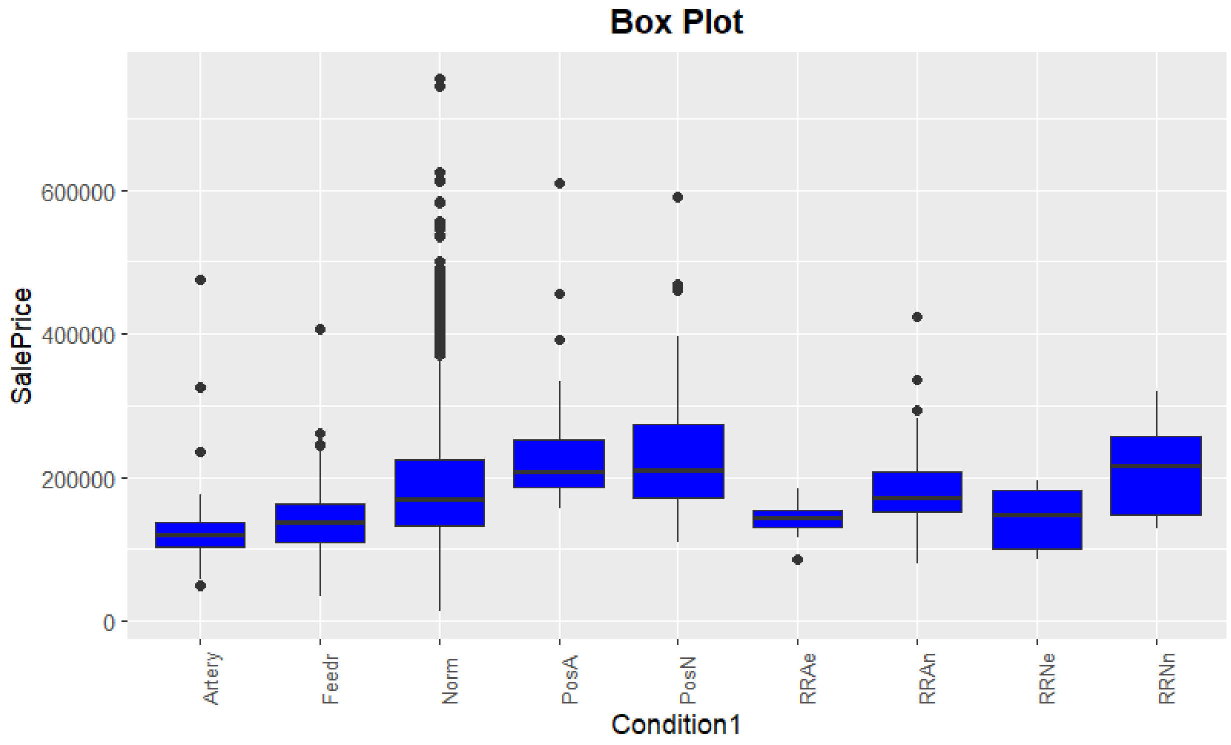
We included the following seven qualitative/categorical properties as part of our top-twenty variables (definitions are from the Ames Housing data documentation):

- KitchenQual - Kitchen quality
- ExterQual - Evaluates the quality of the material on the exterior
- BsmtQual - Evaluates the height of the basement
- HeatingQC - Heating quality and condition
- Condition1 - Proximity to various conditions
- Condition2 - Proximity to various conditions (if more than one is present)
- Neighborhood - Physical locations within Ames city limits

For quality-based ratings above such as KitchenQual, ExterQual, BsmtQual, and HeatingQC, while most homes are rated typical/average (‘TA’) or good (‘Gd’), homes with better ratings (‘Gd’ and excellent ‘Ex’) yield higher SalePrice. A box plot distribution of ExterQual to SalePrice is shown below.



In regard to Condition1 and Condition2, homes that are near or adjacent to positive off-site features (‘PosN’ or ‘PosA’ respectively) such as a park, yield higher SalePrice. A box plot distribution of Condition1 to SalePrice is shown below.



We performed data cleanup with BsmtQual as it had null values. Based on the Ames Housing data dictionary, we replaced null values with ‘NA’ as it denotes homes with no basement.

While we will go into deeper detail regarding Neighborhood in the next section below, at first glance certain neighborhoods yield higher SalePrice.

3. Initial Exploratory Data Analysis (EDA)

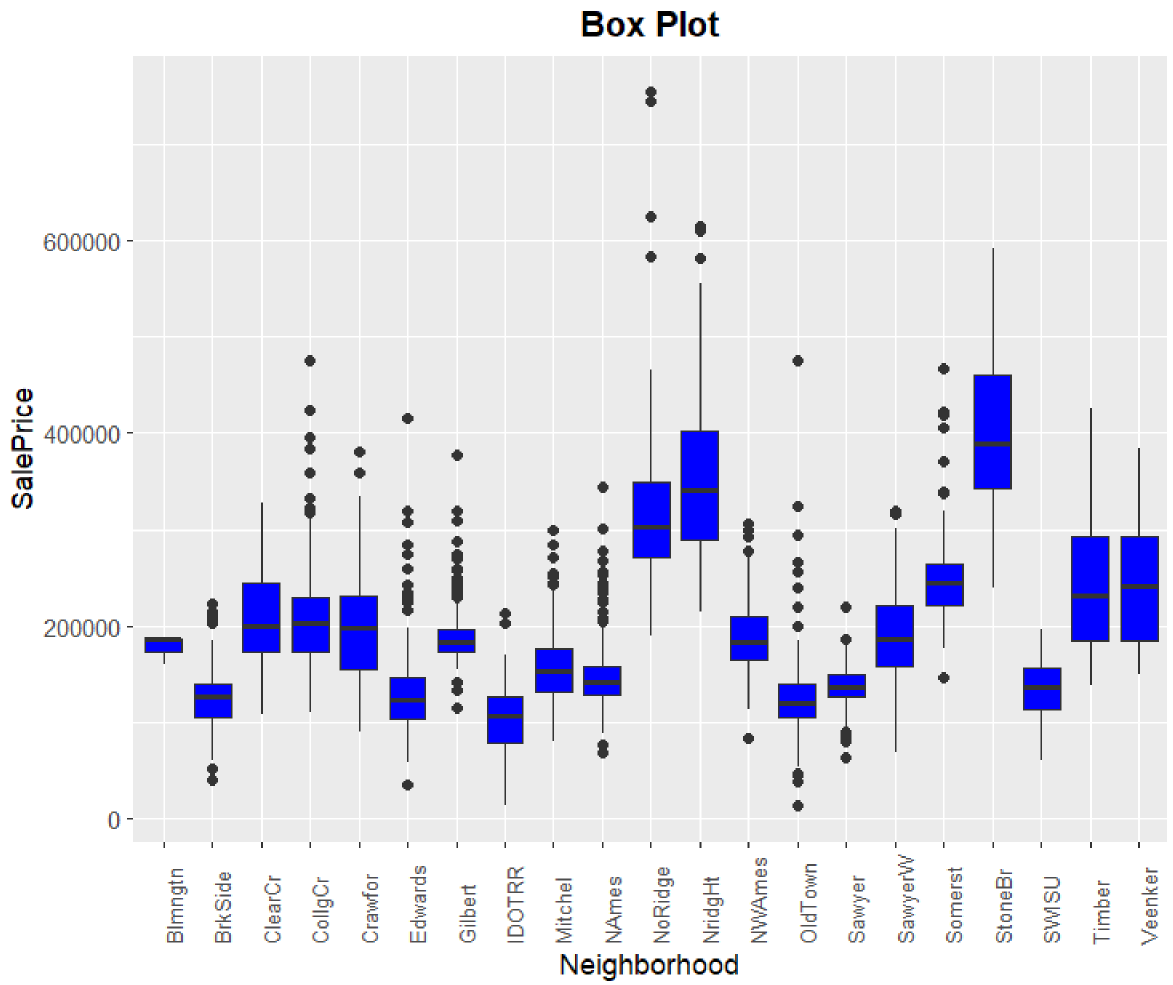
We will further narrow down to the following ten variables:

- Neighborhood
- TotalFloorSF
- TotalBsmtSF
- TotRmsAbvGrd
- FullBath
- GarageCars
- LotArea
- YearBuilt
- MasVnrArea
- QualityIndex

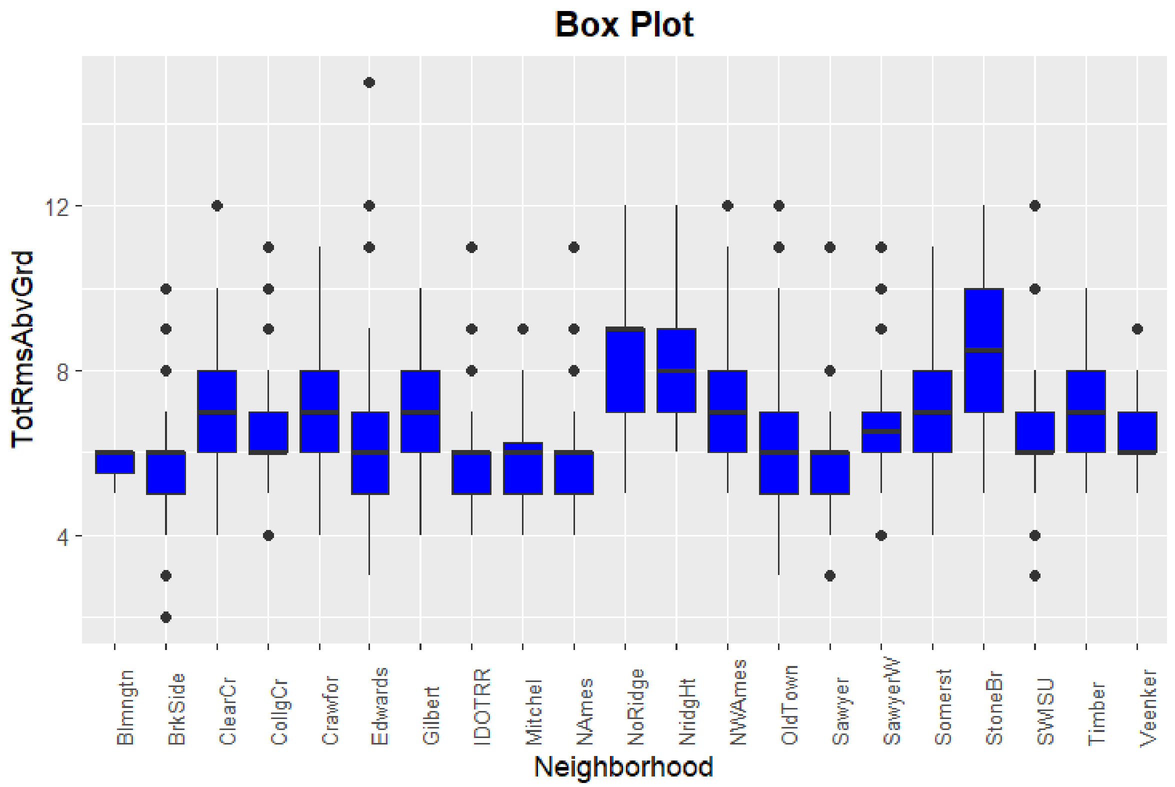
We will perform an exploratory data analysis (EDA) on the above features.

Neighborhood

Neighborhood is the only categorical/qualitative feature to be included in our top ten variables. Based on the box plot below, certain neighborhoods yield higher SalePrice:



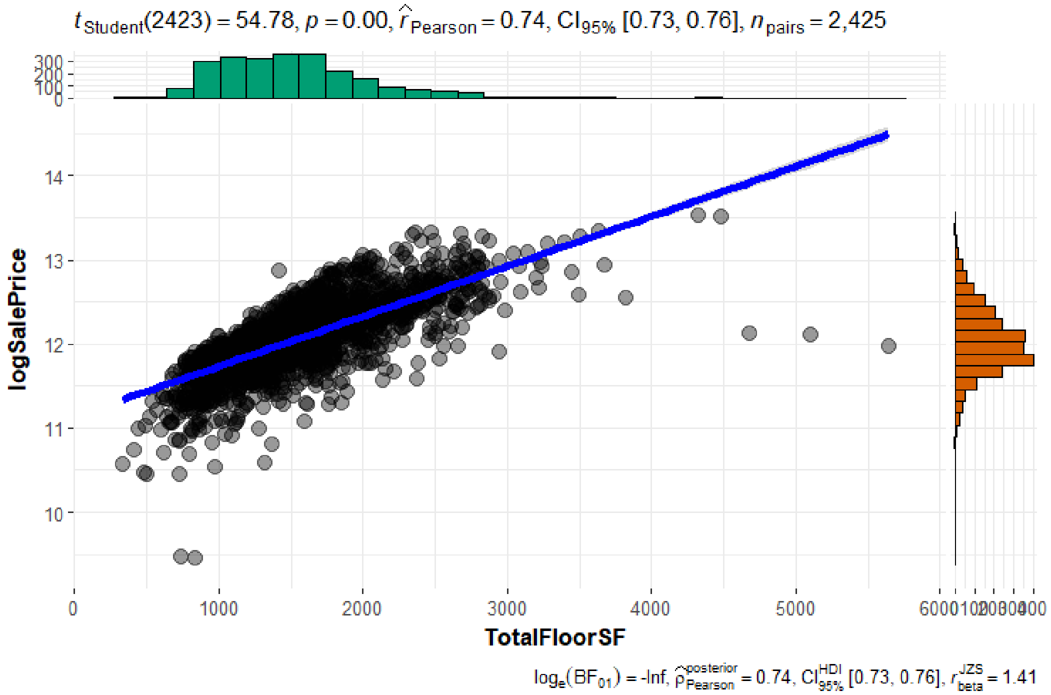
In particular we see Northridge (‘NoRidge’), Northridge Heights (‘NridgHt’) and Stone Brook (‘StoneBr’) neighborhoods yield the highest SalePrice compared to other neighborhoods in Ames. We also find that homes in these neighborhoods are larger as well, noted by total flooring by square feet (‘TotalFloorSF’):



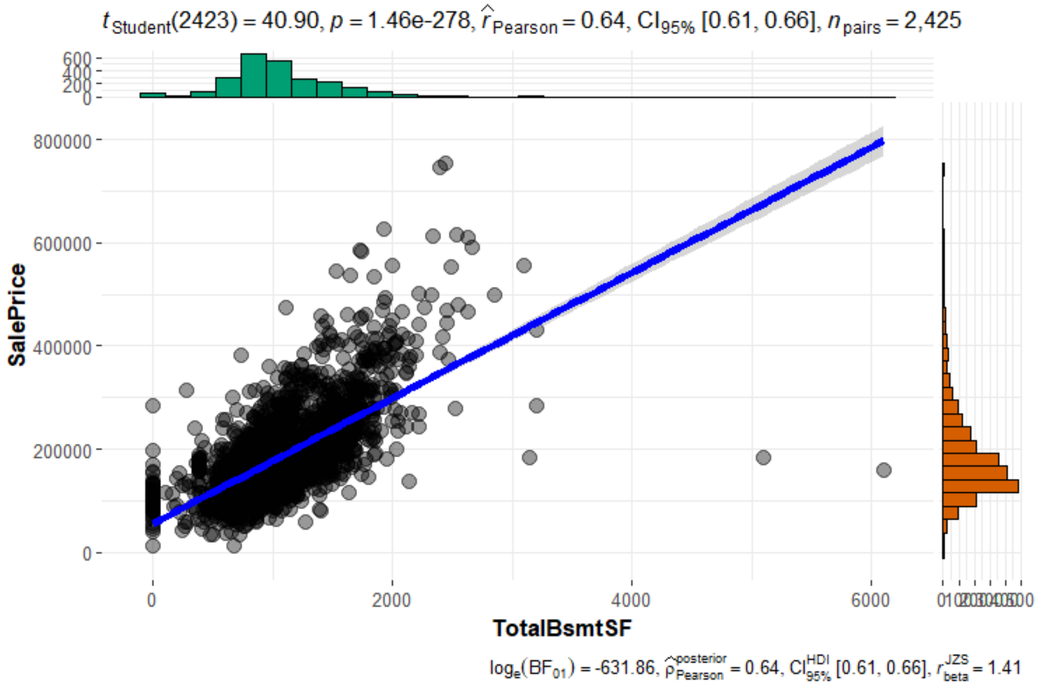
We might infer that larger homes can yield higher SalePrice based on our observations with Neighborhood as it correlates to SalePrice. We'll explore that next.

TotalFloorSF, TotalBsmtSF, TotRmsAbvGrd, FullBath, and GarageCars: Size Matters

We can see a high correlation between TotalFloorSF and SalePrice, as shown in the scatterplot below.

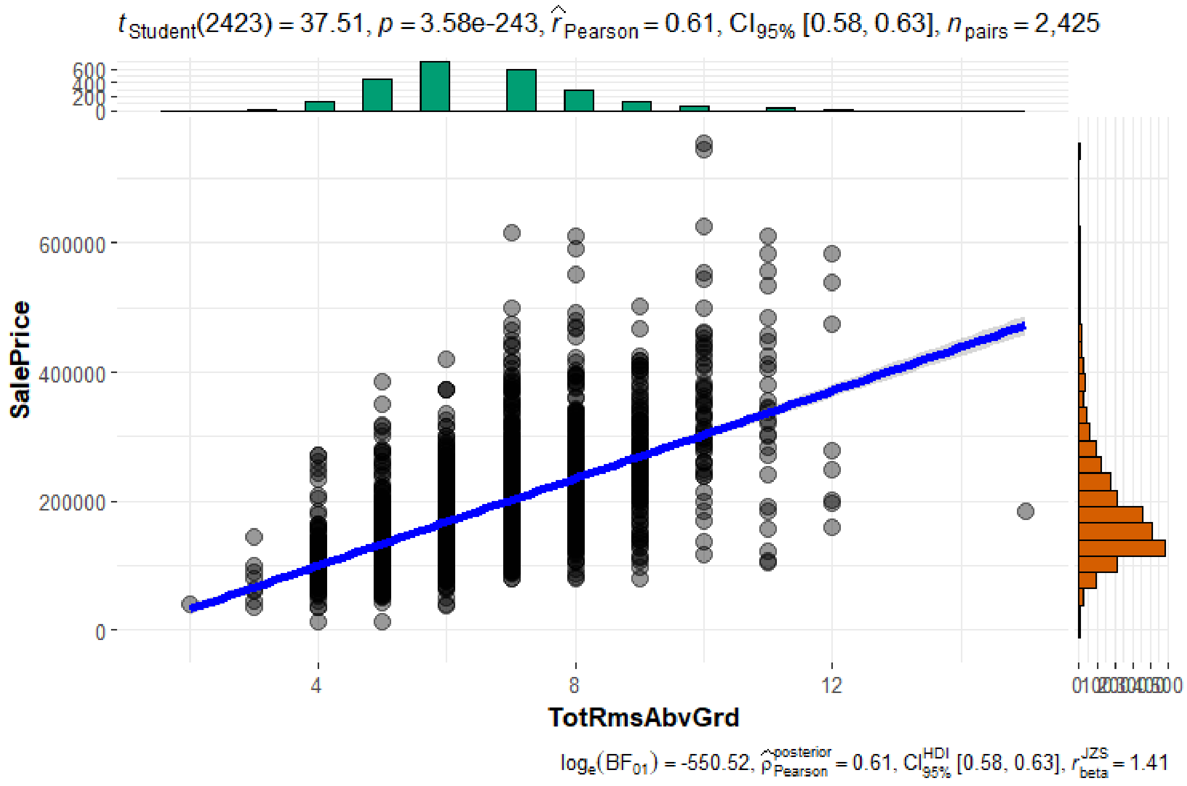


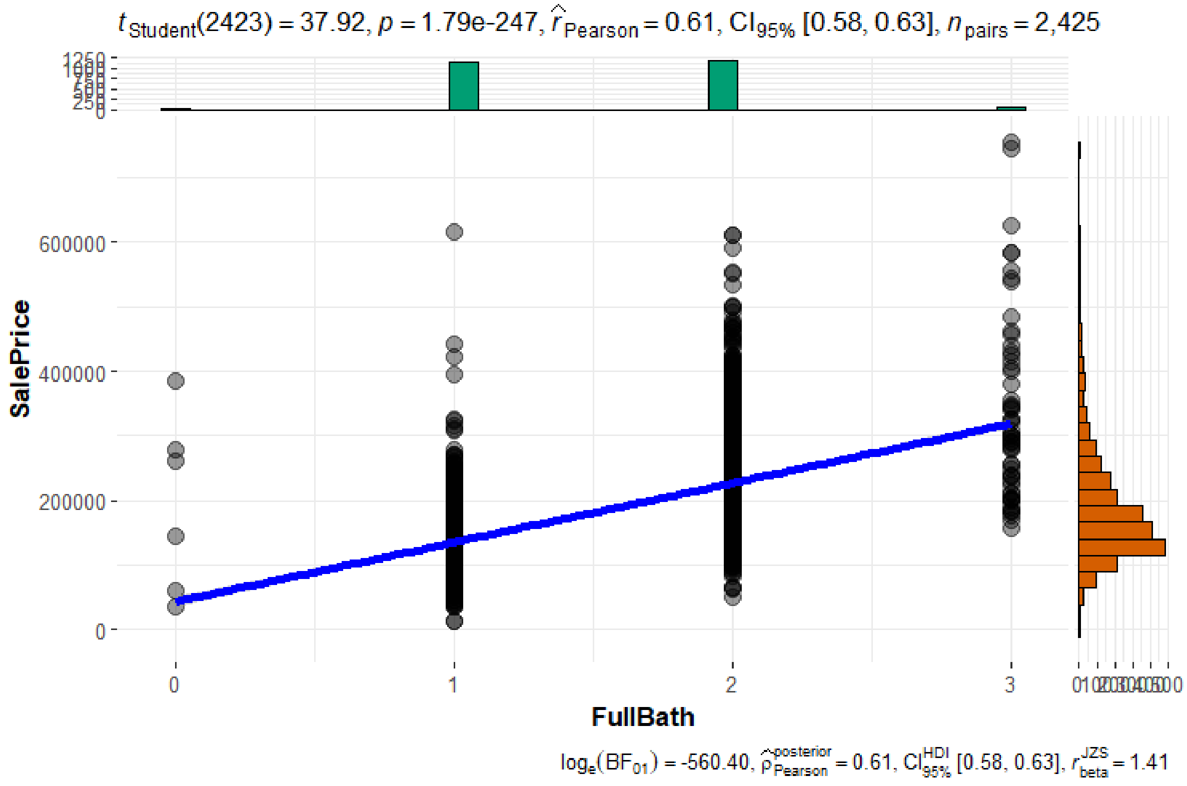
TotalFloorSF only includes areas in a home that are above ground. Many homes in Ames, Iowa also include a basement, which also correlates well with SalePrice, as shown in the scatterplot below.

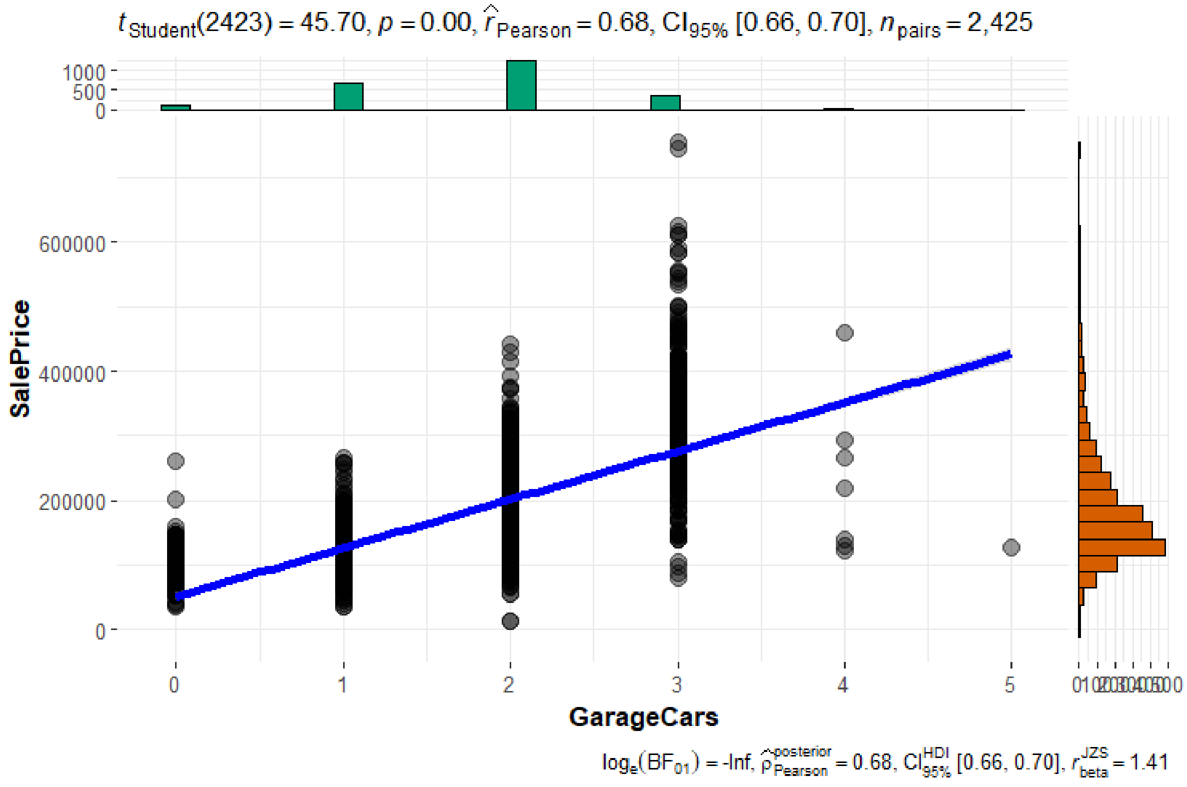


From the chart above, we see that homes that do not include a basement (TotalBsmtSF = 0) seem to yield lower SalePrice.

We can also see a positive correlation between the number of rooms in a home (‘TotRmsAbvGrd’) to SalePrice, as with the number of full bathrooms (‘FullBath’) and the number of cars in the garage (‘GarageCars’):




$t_{\mathrm{Student}}(2423) = 37.92$, $p = 1.79\text{e-}247$, $\widehat{r}_{\mathrm{Pearson}} = 0.61$, $\mathrm{CI}_{95\%}$ [0.58, 0.63], $n_{\mathrm{pairs}} = 2{,}425$
1250
1000
750
500
250
0
600000
400000
200000
0
SalePrice
0
1
2
3
FullBath
0
100
200
300
400
500
$\log_e(\mathrm{BF}_{01}) = -560.40$, $\widehat{\rho}^{\,\mathrm{posterior}}_{\mathrm{Pearson}} = 0.61$, $\mathrm{CI}^{\mathrm{HDI}}_{95\%}$ [0.58, 0.63], $r^{\mathrm{JZS}}_{\mathrm{beta}} = 1.41$

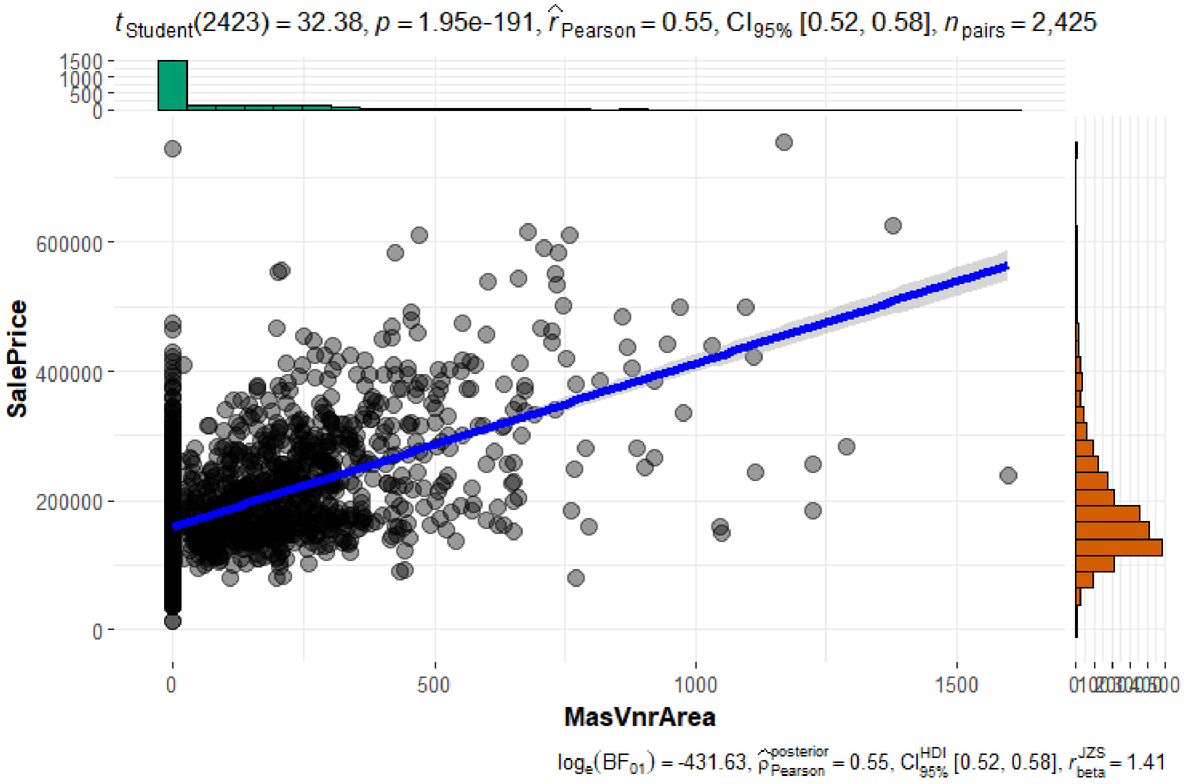

$t_{\mathrm{Student}}(2423) = 45.70$, $p = 0.00$, $\widehat{r}_{\mathrm{Pearson}} = 0.68$, $\mathrm{CI}_{95\%}$ [0.66, 0.70], $n_{\mathrm{pairs}} = 2{,}425$
1000
500
0
600000
400000
200000
0
SalePrice
0
1
2
3
4
5
GarageCars
0
100
200
300
400
500
$\log_e(\mathrm{BF}_{01}) = -\mathrm{Inf}$, $\widehat{\rho}^{\,\mathrm{posterior}}_{\mathrm{Pearson}} = 0.68$, $\mathrm{CI}^{\mathrm{HDI}}_{95\%}$ [0.66, 0.70], $r^{\mathrm{JZS}}_{\mathrm{beta}} = 1.41$

Similar to basements, homes without a garage (GarageCars = 0) yield a lower SalePrice. There are some exceptional outliers with very large 4- or 5-car garages that do not yield a high SalePrice, but many 3-car garages yield a higher SalePrice.

Overall, we can see that the larger a home based on the factors above can yield a higher SalePrice. Many of those larger homes are in the Northridge, Northridge Heights, and Stone Brook neighborhoods. It would be interesting to see if these particular neighborhoods could also have a higher household income, but that is beyond the scope of this modeling exercise.

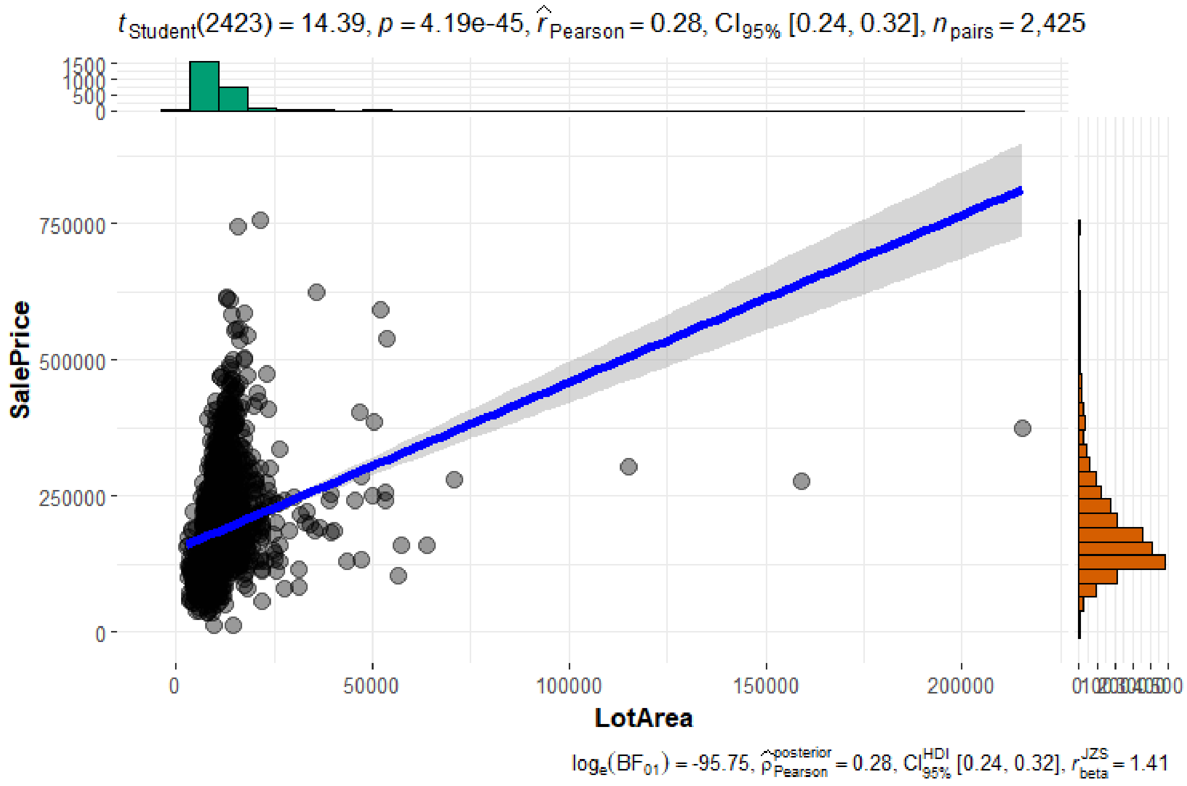
MasVnrArea

There seems to be a good correlation between masonry veneer area ('MasVnrArea') and SalePrice, as shown in the scatterplot below.



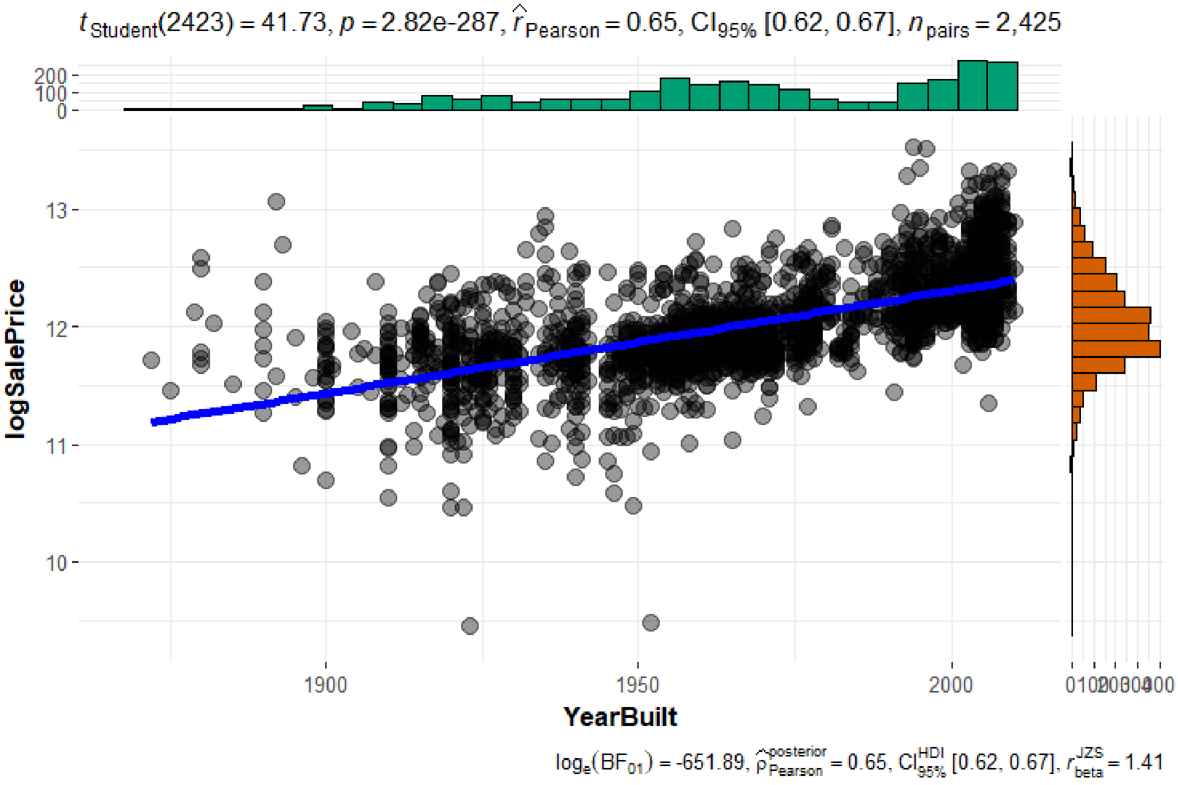
We can also see that the area distribution denoted by the green histogram in the chart above also shows that most homes in the Ames area do not have masonry veneer (MasVnrArea = 0).

LotArea



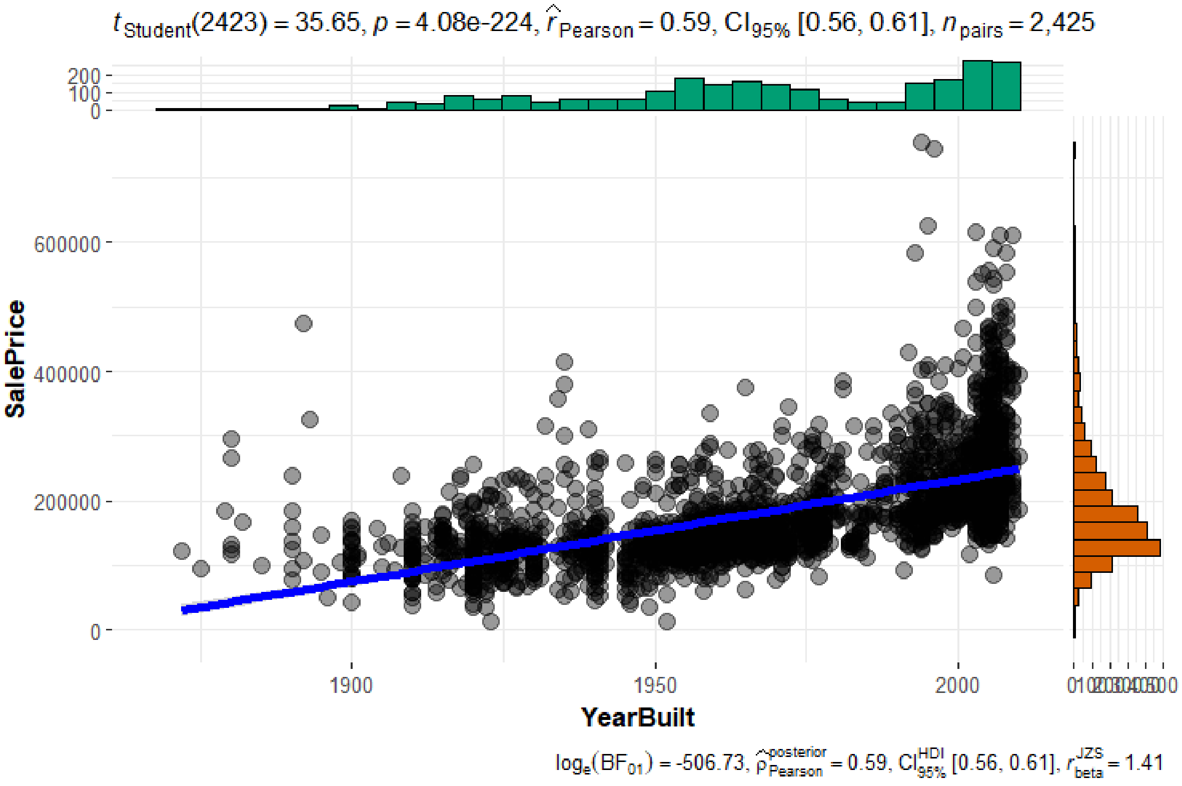
Based on the scatterplot above, we can see that a few outlier homes with very large lots can significantly skew the distribution. If we consider using LotArea in creating a model, we might have to take into account how to handle outliers and see if doing so can improve the correlation with SalePrice.

YearBuilt



$t_{\text{Student}}(2423) = 41.73$, $p = 2.82\text{e-}287$, $\widehat{r}_{\text{Pearson}} = 0.65$, $\text{CI}_{95\%}$ [0.62, 0.67], $n_{\text{pairs}} = 2{,}425$

$\log_e(\text{BF}_{01}) = -651.89$, $\widehat{\rho}_{\text{Pearson}}^{\,\text{posterior}} = 0.65$, $\text{CI}_{95\%}^{\text{HDI}}$ [0.62, 0.67], $r_{\text{beta}}^{\text{JZS}} = 1.41$



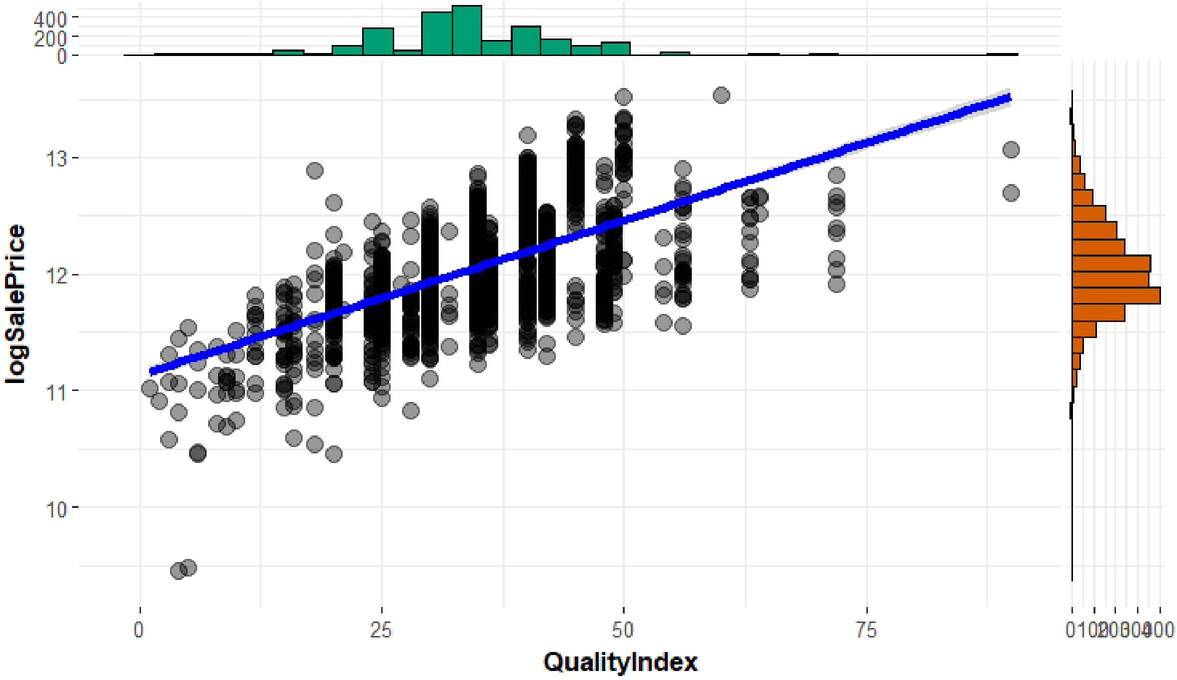
$t_{\text{Student}}(2423) = 35.65$, $p = 4.08\text{e-}224$, $\widehat{r}_{\text{Pearson}} = 0.59$, $\text{CI}_{95\%}$ [0.56, 0.61], $n_{\text{pairs}} = 2{,}425$

$\log_e(\text{BF}_{01}) = -506.73$, $\widehat{\rho}_{\text{Pearson}}^{\,\text{posterior}} = 0.59$, $\text{CI}_{95\%}^{\text{HDI}}$ [0.56, 0.61], $r_{\text{beta}}^{\text{JZS}} = 1.41$

The year built variable (‘YearBuilt’) has a stronger positive correlation with the better-distributed spread of log(SalePrice). The YearBuilt distribution shows periods of home building growth during 1950 to 1975, and later growth in the late 90s and in the 2000s. Both scatterplots above show that newer homes yield a higher SalePrice.

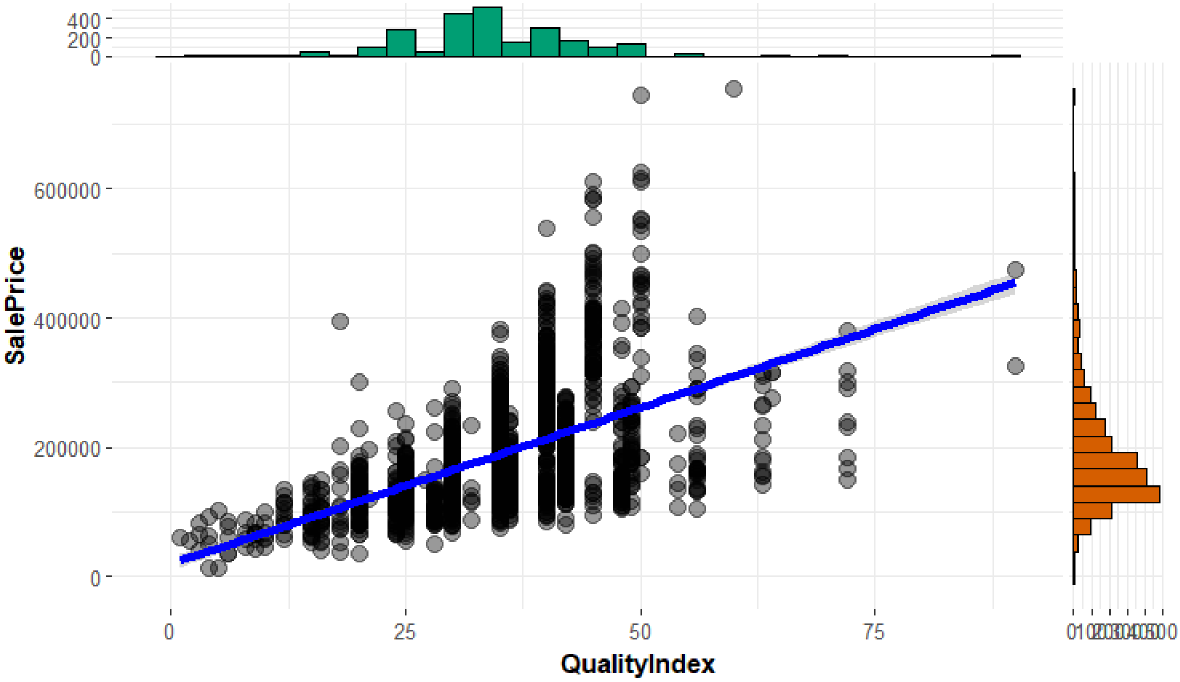
QualityIndex

$t_{\text{Student}}(2423) = 36.84$, $p = 2.36\text{e-}236$, $\widehat{r}_{\text{Pearson}} = 0.60$, $\text{CI}_{95\%}$ [0.57, 0.62], $n_{\text{pairs}} = 2{,}425$



$\log_e(\text{BF}_{01}) = -534.85$, $\widehat{\rho}^{\text{posterior}}_{\text{Pearson}} = 0.60$, $\text{CI}^{\text{HDI}}_{95\%}$ [0.57, 0.62], $r^{\text{JZS}}_{\text{beta}} = 1.41$

$t_{\text{Student}}(2423) = 32.17$, $p = 2.12\text{e-}189$, $\widehat{r}_{\text{Pearson}} = 0.55$, $\text{CI}_{95\%}$ [0.52, 0.57], $n_{\text{pairs}} = 2{,}425$



$\log_e(\text{BF}_{01}) = -426.95$, $\widehat{\rho}^{\text{posterior}}_{\text{Pearson}} = 0.55$, $\text{CI}^{\text{HDI}}_{95\%}$ [0.52, 0.57], $r^{\text{JZS}}_{\text{beta}} = 1.41$

The quality index (‘QualityIndex’) is a data transformation multiplying the values of overall quality(‘OverallQual’) and overall condition (‘OverallCond’) from the Ames dataset. Based on the data dictionary of the original properties, the distribution of the index shows most homes lying in the above average to good range. The quality index has a stronger positive correlation with log(SalePrice).

4. EDA for Modelling

The top three properties from the Ames dataset we should use for any model to predict SalePrice are:

- Neighborhood
- TotalFloorSF
- QualityIndex

As many realtors would say, ‘Location! Location! Location!’ The box plot of Neighborhood to SalePrice shown earlier shows how certain neighborhoods yield higher SalePrice. In those pricier neighborhoods we also see a correlation to larger homes, and that’s why the transformed TotalFloorSF variable should be included and is highly correlated. The quality index might not have as high correlation to other properties such as number of rooms, bathrooms, and garage cars, it has a much broader range of values and takes into account the mix of both overall quality and condition. Comparing the property against log(SalePrice) gives it a greater correlation as well. While few homes in the Ames dataset are rated ‘Excellent’ in overall condition or quality, many are still rated as ‘Above Average and Good’ and in the upper half of the criteria.

5. Summary/Conclusions

The initial twenty variables chosen provide a very good foundation for creating predictive models for the price of a home in Ames, Iowa, many based on their strong correlation to SalePrice. Going forward, quality-based categories such as KitchenQual, ExterQual, BsmtQual, and HeatingQC can be transformed into ranked quantitative variables without much difficulty. Condition-based variables can possibly be transformed into ranked quantitative variables as well. Ranking neighborhoods based on their strong correlation to SalePrice, though, could be problematic. Much care and further research should be considered when attempting to transform and quantify neighborhoods, as it can lead to issues of systemic bias and income inequality. We found that TotalFloorSF and QualityIndex are useful transformations to consider for modeling due to their strong correlations to SalePrice and log(SalePrice). They are just our initial transformations and as we further explore the Ames dataset there could possibly be more opportunities for other kinds of data transformations along the way as needed.