Capstone 1

Home prices in Ames, Iowa

Introduction

- Pricing a home for sale is a challenge all real estate agencies face daily.
- The current process relies on manually comparing a home
- The process entails comparing a home to other similar homes in sold in the area
- It is labor intensive and error prone

The dataset used can be found below:

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Includes **79** home features with over **1459** rows of data

Initial Assumptions

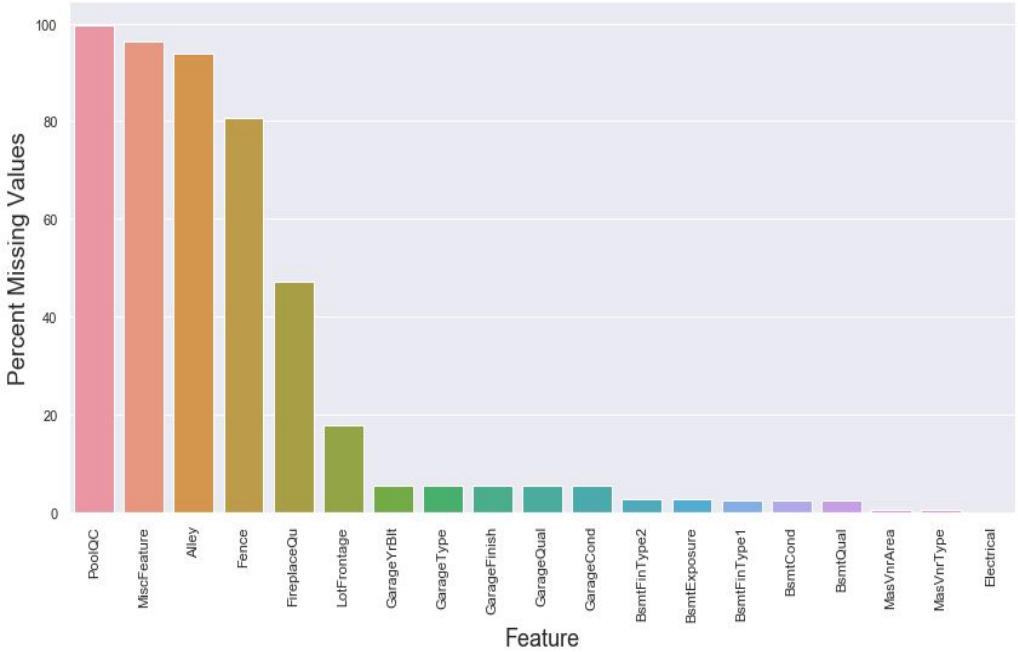
- Location, Location, Location! - Location within Ames, Iowa will be critical in determining sale price.
- Size of the home will also play a large role - There are many variables measured in square footage
- Home sales typically fluctuate throughout the year. Could this pattern be confirmed visually? If this pattern is present, could it mean that sales price may be affected by these fluctuations?

Let’s move on to some data wrangling and cleaning in order to prepare for exploration of these assumptions.

Data Wrangling Steps

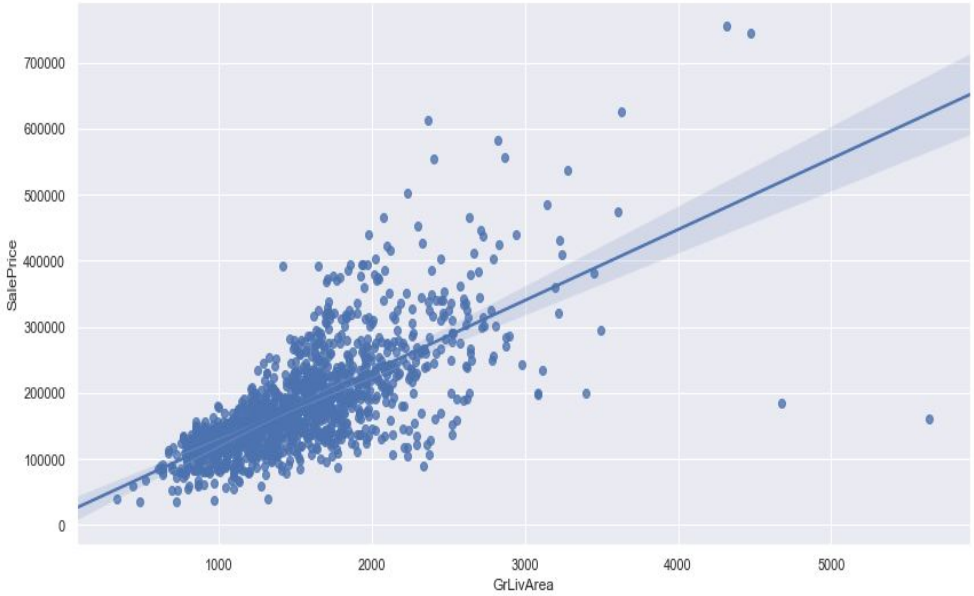
Percent missing data

- Many null values can be replaced with a more logical value.
- For example, **GarageType**, **GarageQual** and **GarageCond** missing data is described as meaning ‘No Garage’ in the data dictionary
- **BsmntQual**, **BsmtCond** null values indicate ‘No Basement’



Dealing with Outliers

- Dataset author mentions there are 4 outliers in the dataset
- Specifically anything greater than 4000 square feet living area should be considered as an outlier
- However, before elimination it makes sense to look at these outliers visually and understand a bit more about them
- The four data points in question can be identified fairly clearly in this visualization. Lets isolate these data points.



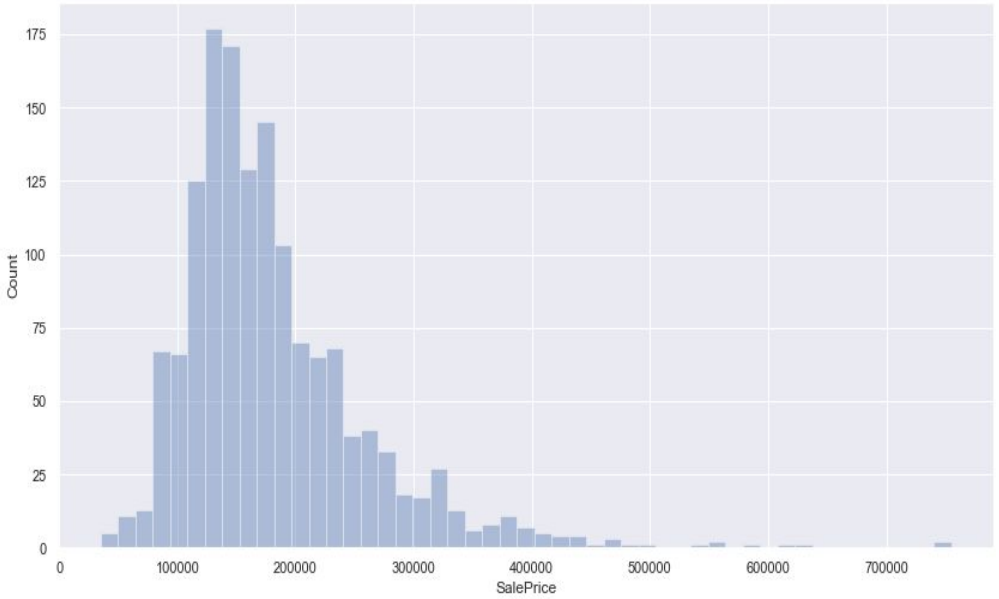
Dealing with outliers, cont.

- Index 691 and 1182 could be realistic data points
- They were sold at a very high price point
- For now all outliers will be maintained in the dataset
- For modeling purposes we don't want to assume a purely linear relationship, even if appearances show otherwise

Index	SalePrice	GrLivArea
523	184750	4676
691	755000	4316
1182	745000	4476
1298	160000	5642

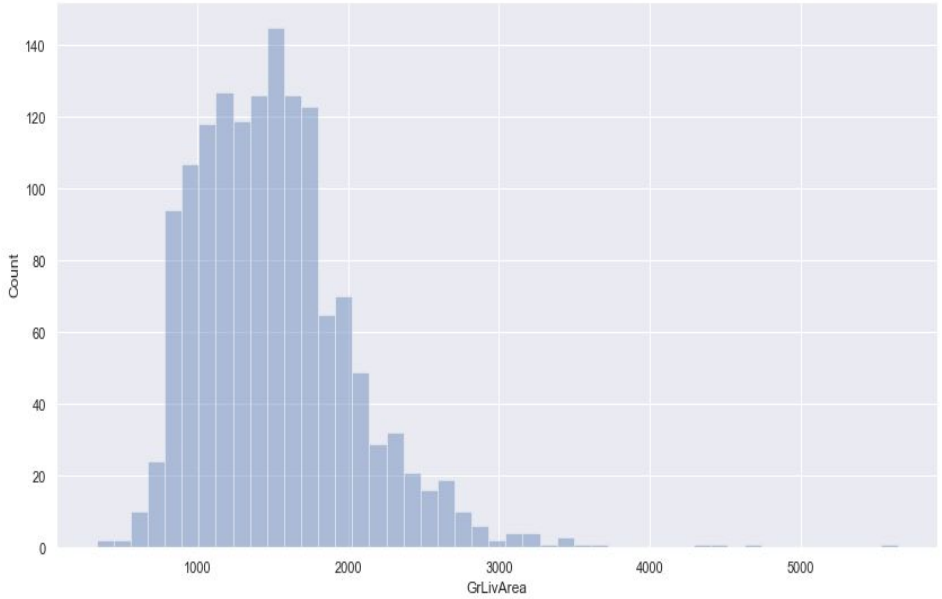
Sale Price Distribution

- The distribution seems skewed to the right, with what look like several outliers
- Let’s take a look at what seems to be the most obvious influencers on sales price, livable square footage (the variable GrLivArea)



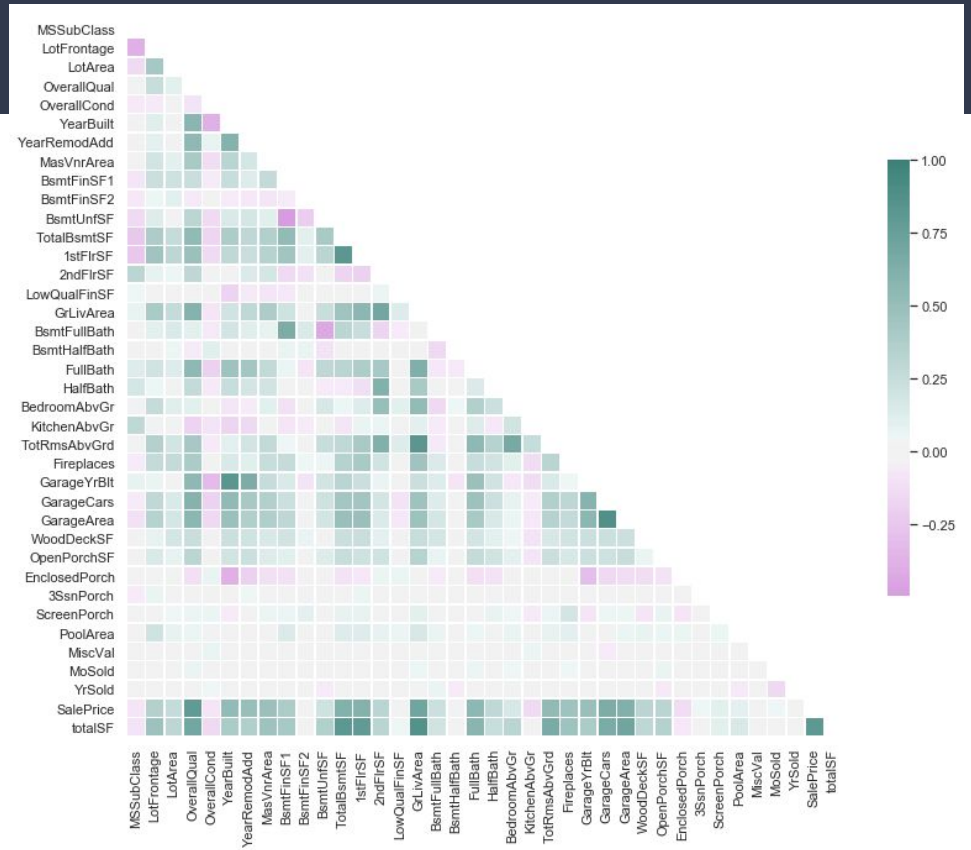
GrLivArea (Greater Living Area) Distribution

- The distribution is very similar to that of sale price itself.
- This is further evidence that this feature is a good predictor for sale price.
- Also a justification of the right skew of the sale price distribution



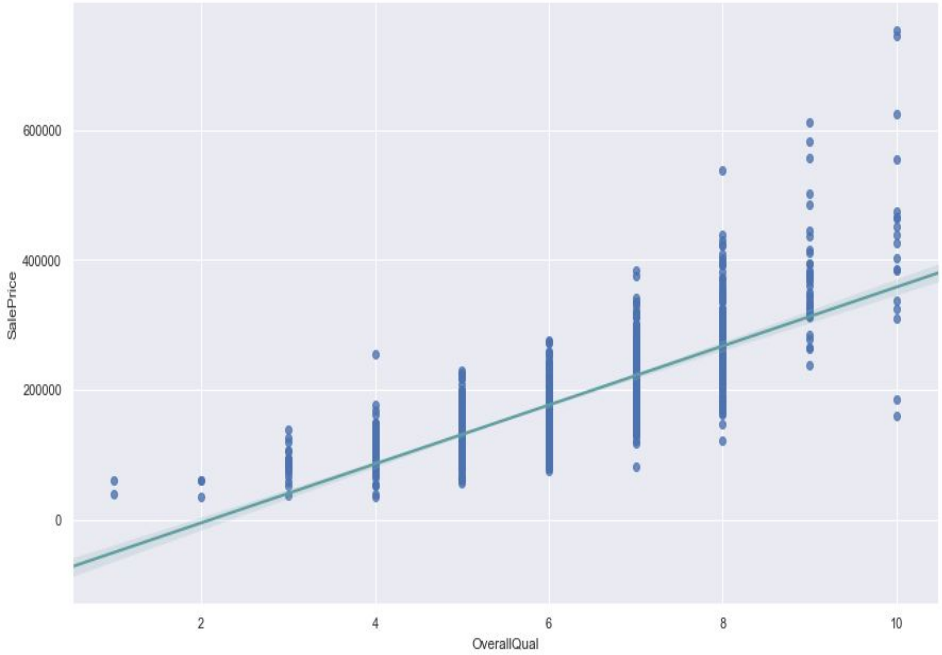
Numerical Correlations to Sale Price

- Correlation matrix to the right is a heatmap indicating the pearson correlation coefficients for each of the variables
- This visual excludes categorical data and only looks at numerical types.
- Each of the numerical variables are compared to one another until all of their correlation coefficients are computed
- The result shows that **OverallQual** has the highest correlation to **SalePrice** followed by several other square footage related parameters.



OverallQual

- Since its the highest correlation reported, it's worthwhile to explore it visually
- The score ranges from 1-10 and is evaluated per home.
- Seems there is a correlation here
- Notice the wide range of prices for each quality score

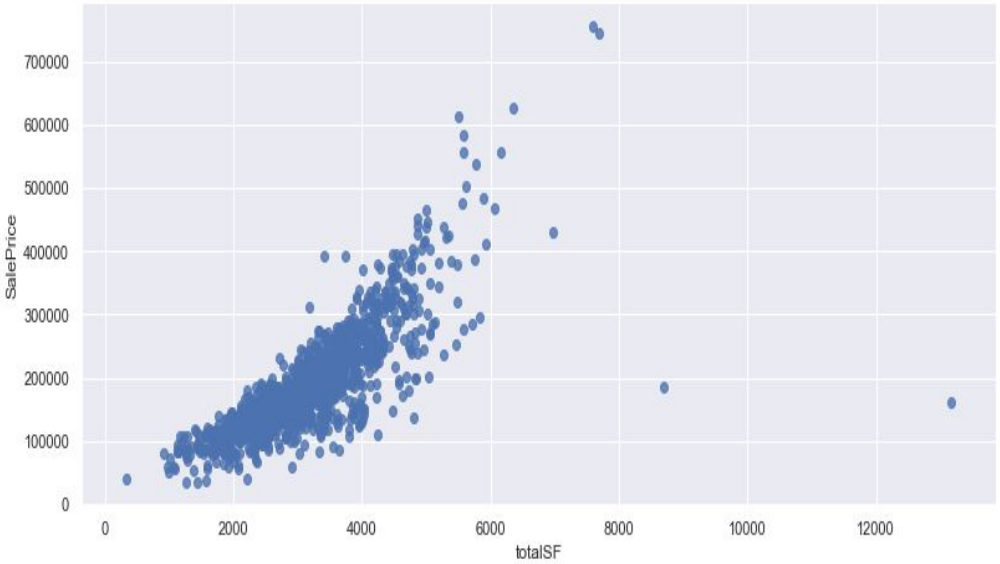


Total Square Footage of Homes

- Many square footage parameters showed correlation to SalePrice
- It makes sense to combine some of these where possible
- We can compute a new parameter **totalSF** with the following expression:

totalSF = GrLivArea + TotalBsmtSF + GarageArea

New totalSF parameter is plotted against SalePrice:



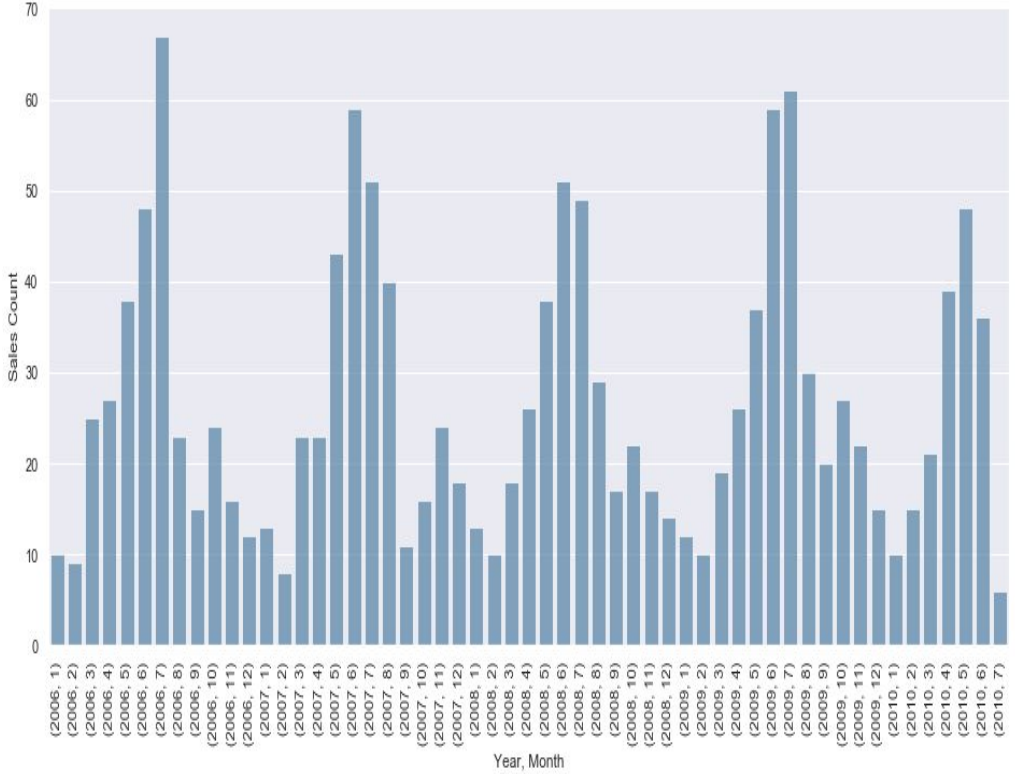
Pearson correlation for the new totalSF parameter

- Correlation between totalSF and SalePrice seems very strong
- When we evaluate the Pearson coefficient we obtain a result of ~ .807 which now tops the charts.
- Previously the strongest correlation was calculated as **OverallQual**

$$P_{totalSF}(.807) > P_{OverallQual}(.791)$$

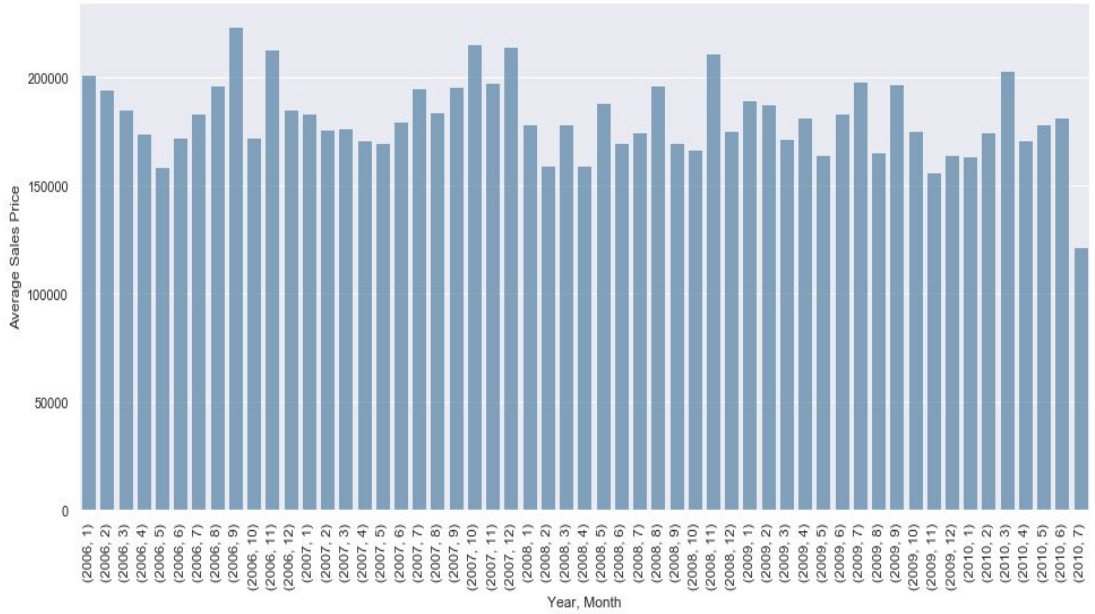
Evaluating Sales Volume

- Sales volume can be defined as the number of homes sold in a given period
- To evaluate this we can combine the month sold (**MoSold**) and year sold (**YrSold**) parameters
- We then aggregate entries to create a count per month/year period
- The result shows a cyclical pattern to sales volume
- Peaks can be observed in summer months with troughs in the winter



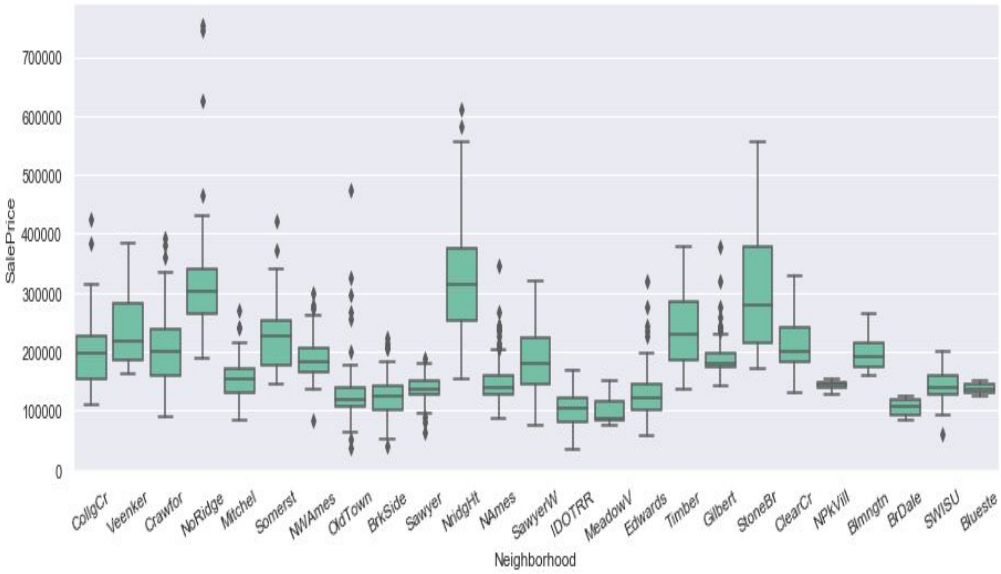
Average sales price - Per month, year period

- There is a trend observable, but doesn't seem to be as strong
- This may not be the strongest predictor of sale price



Location, Location, Location

- Aggregate sales price by neighborhood to produce a categorical box plot
- Significant variance can be observed between the distributions
- Seems indicative that neighborhood plays a significant role in predicting Ames, Iowa sale prices
- It’s worth further investigation to truly determine if we can confirm this significance statistically

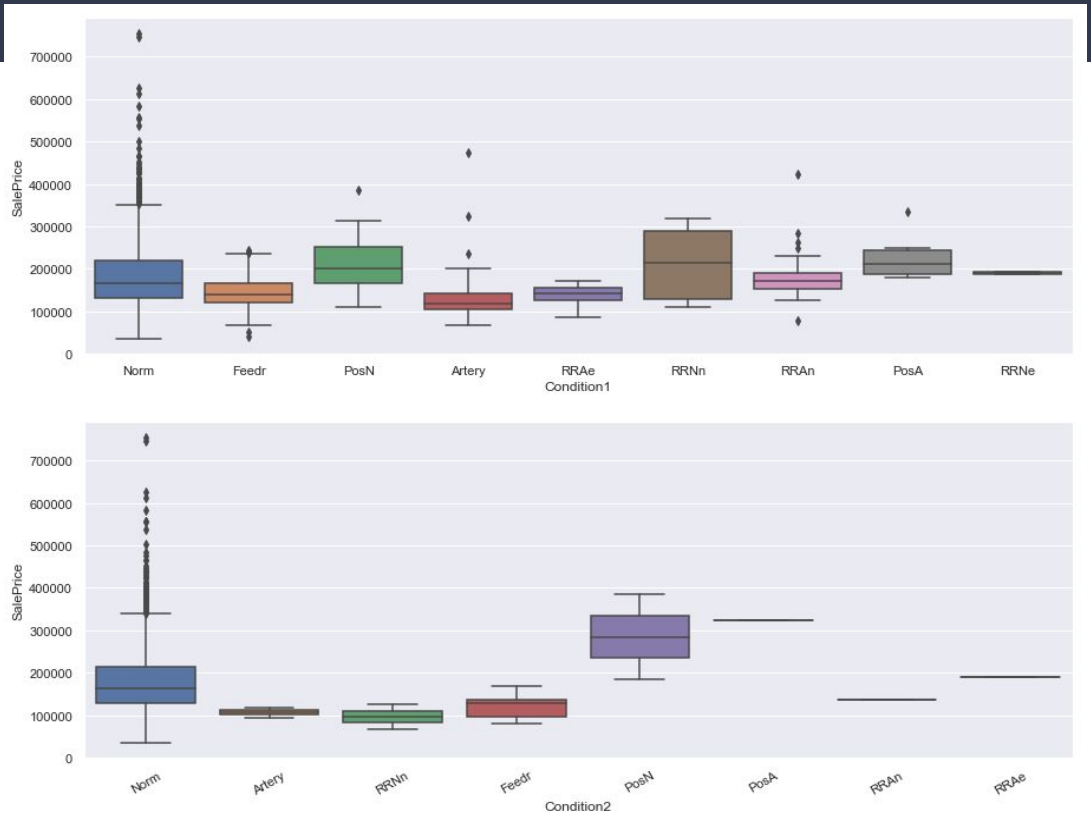


Other Categorical Visualizations

We will take a look at **Condition1** and **Condition2**.

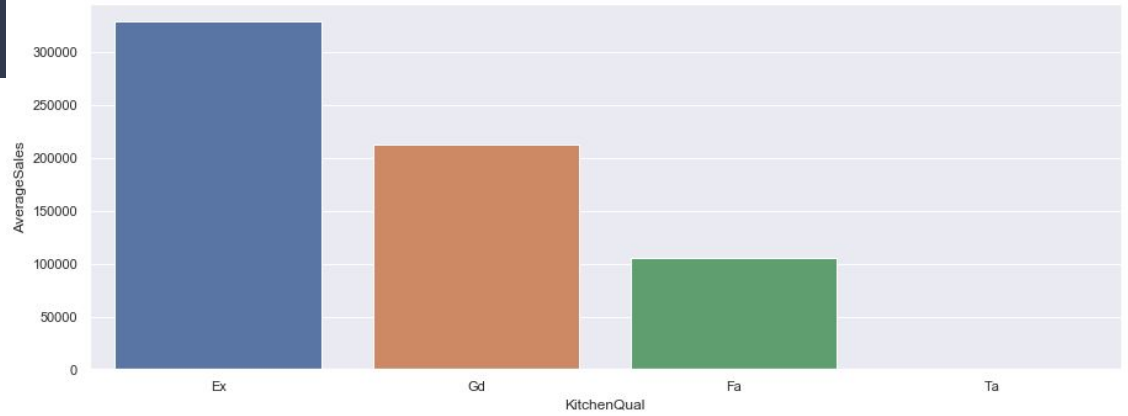
These are features which specify the proximity to certain city conditions

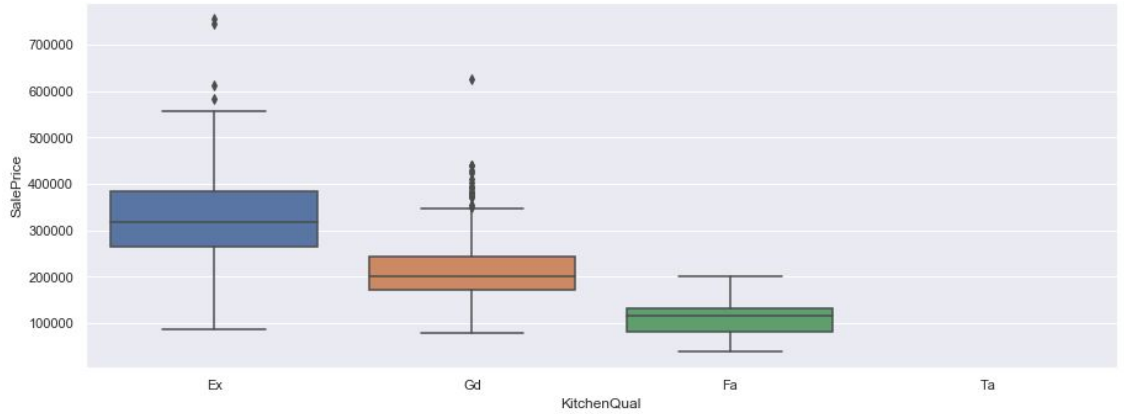
For example: **PosN** indicates a positive off-site feature including a park).



Kitchen Quality

Let’s take a look at kitchen quality denoted by the KitchenQual variable:

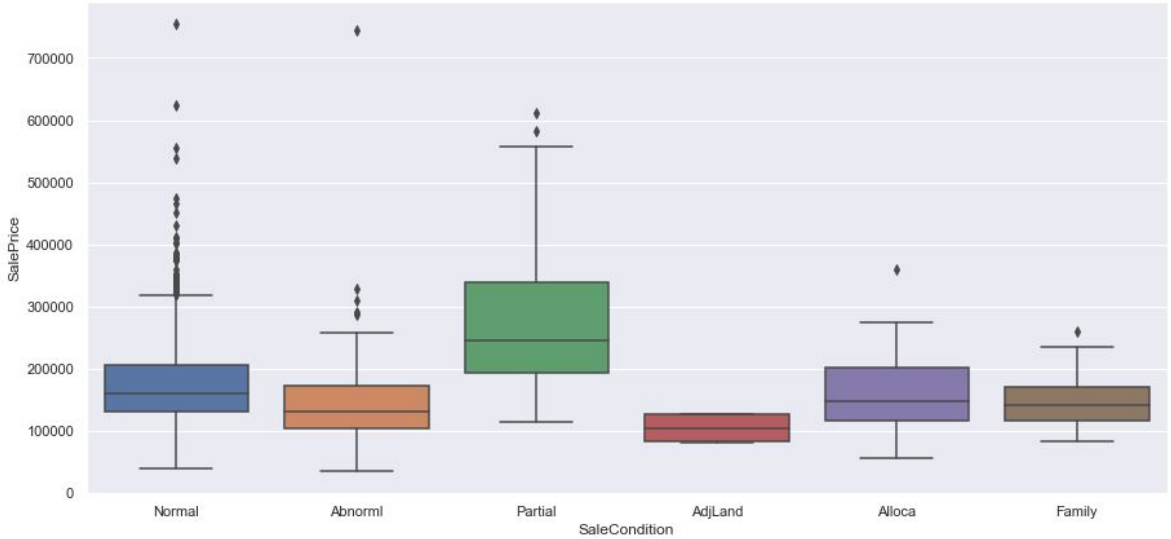




Sale Condition

Let’s also examine the SaleCondition feature.

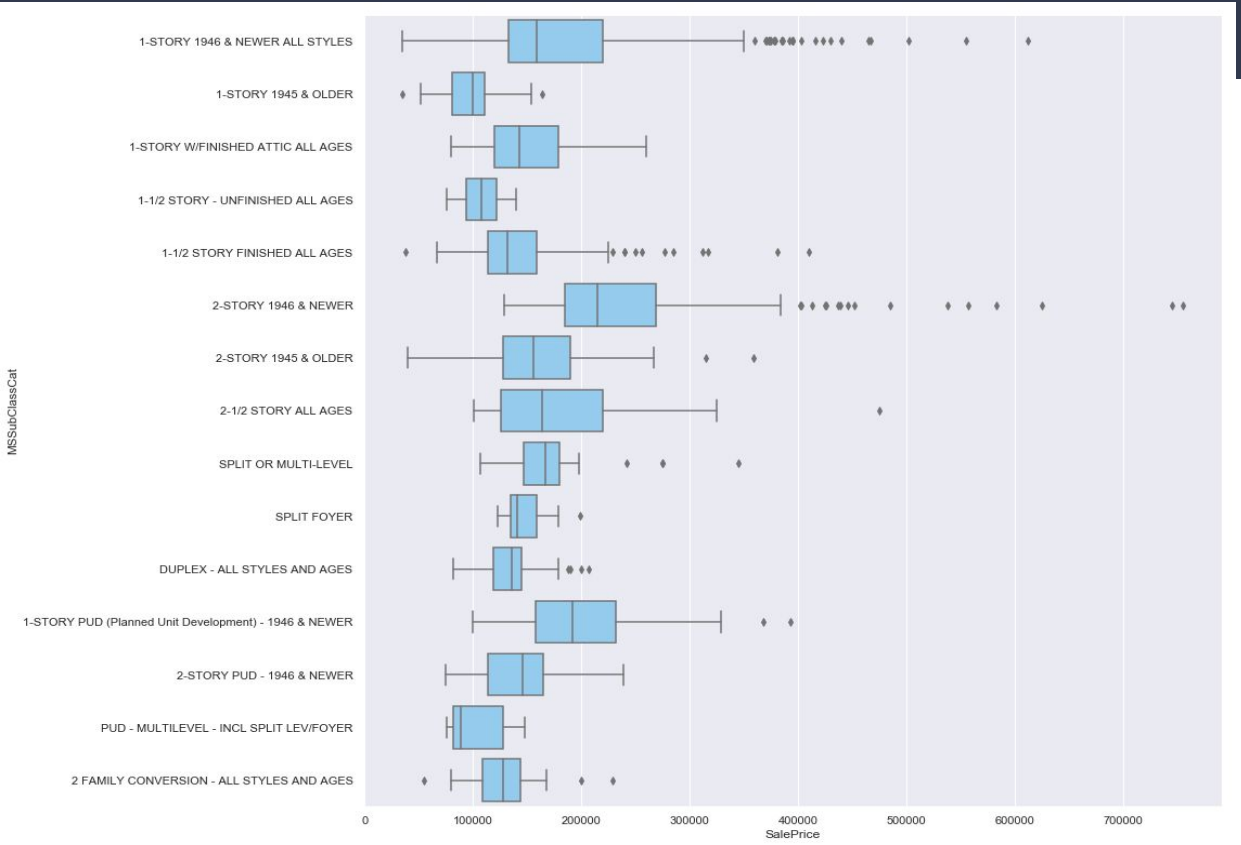
New homes are denoted by the ‘**Partial**’ value.



Type of Dwelling: MSSubClass

The type of dwelling in the sale is establilshed by the **MSSubClass** parameter.

This parameter was translated from a numerical (unintelligible) value to a much clearer descriptive value (as shown on y-axis of plot)



Summary of EDA Results

- Results show that square footage seems to have the strongest correlation to home sale price
- Neighborhood also seems relevant and requires further investigation to establish statistical significance.
- totalSF, the new parameter created by combining square footage parameters, has the strongest Pearson correlation
- Sales volume does not seem to significantly affect home sale price.

Up Next: Some inferential statistics should be applied to determine if difference observed between neighborhoods are truly significant.

The correlation coefficients of totalSF and OverallQual should be compared to officially establish one as the stronger predictor.

Inferential Statistics: Exploring totalSF & OverallQual

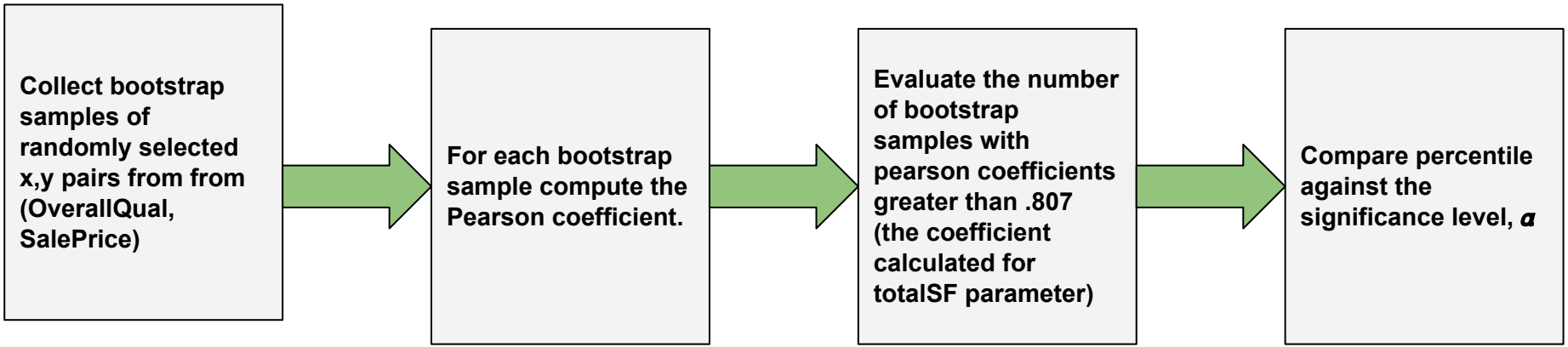
- **totalSF** was calculated to have a Pearson coefficient of .807
- This comes in just above **OverallQual** at .791, so about a .16 difference
- This is worth investigating since it is worth knowing which feature may be a better predictor
- Let's write out the hypothesis as follows:

Ho: The difference in correlation coefficients between **totalSF** and **OverallQual** is negligible.

Ha: The **totalSF** parameter has a greater correlation to **SalePrice**

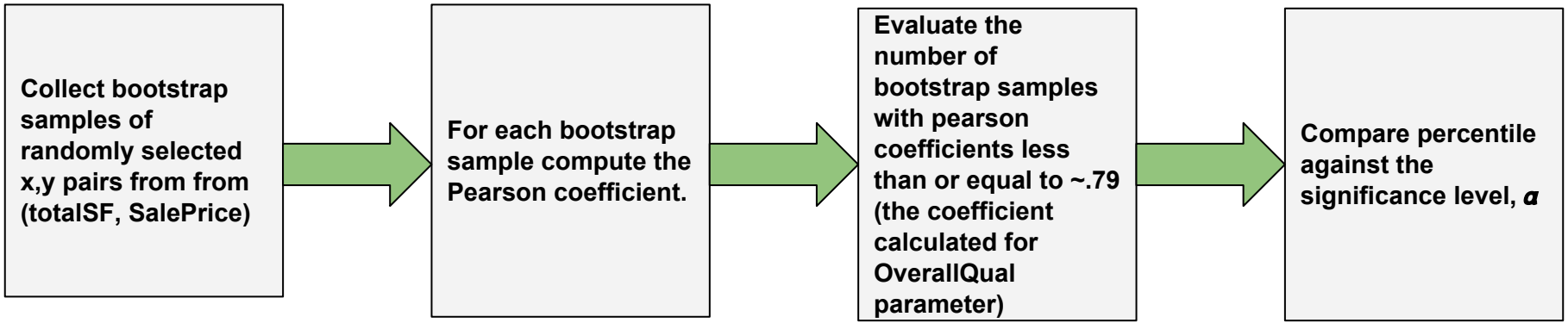
***α*:** .05

Bootstrap Test Approach



Bootstrap Test Results

- With a p-value of .05 we would normally accept our null hypothesis since p>=***α***
- However in this case we are exactly on the threshold so this decision becomes quite difficult.
- What happens when we look at this the other way around?
- Lets approach the problem this way:



Results, Continued

- Results show that 28.86% of the replicates have coefficients calculated to be .79098 or lower.
- Given this result we are led to **accept the null hypothesis** that the difference between correlations is negligible.
- Finally using scipy.stats in python we compute the coefficients and their relative p-values with the following result:

totalSF pearson:

(P= 0.8075184760515013, p-value = 0.0)

OverallQual pearson:

(P=0.7909816005838051, p-value = 2.185675268e-313)

Inferential Statistics: Neighborhood

Let's examine the mean sale price aggregated by **Neighborhood:**

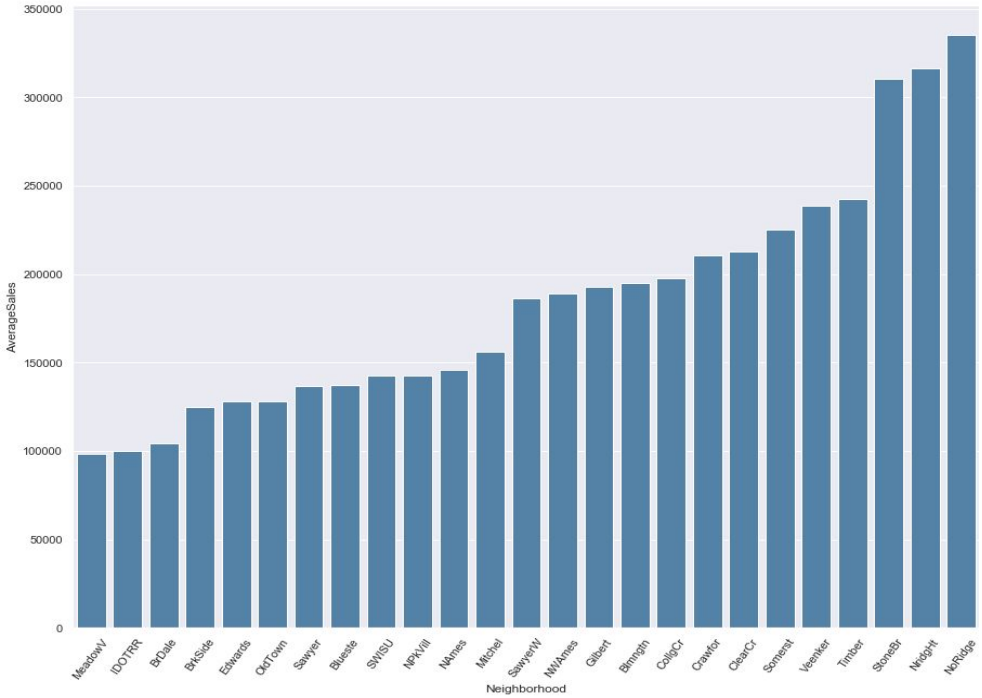
From the barplot we see large variations in mean sale price. How many of these means differ?

Lets answer this question with a hypothesis test:

Ho = There is no difference between mean of MeadowV and Timber home sale prices.

Ha = The means are not equal (two tailed test)

α= Lets use an alpha value of .05 or 5%

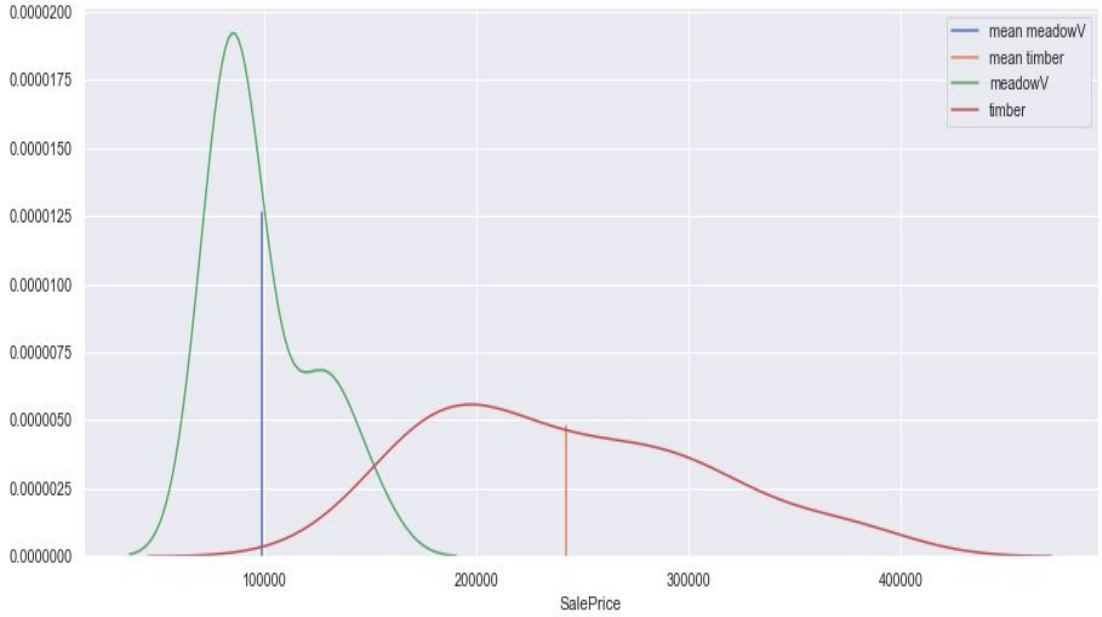


Examining Distribution and Summary Statistics

- **Standard Error Calculated :** 11759.927
- **Z-value Calculated:** -12.217
- **P-value obtained:** 2.52 e-34

From this p-value, its quite clear that the means between these two neighborhoods differ significantly. So we can reject the null hypothesis **Ho**

How about other neighborhoods?

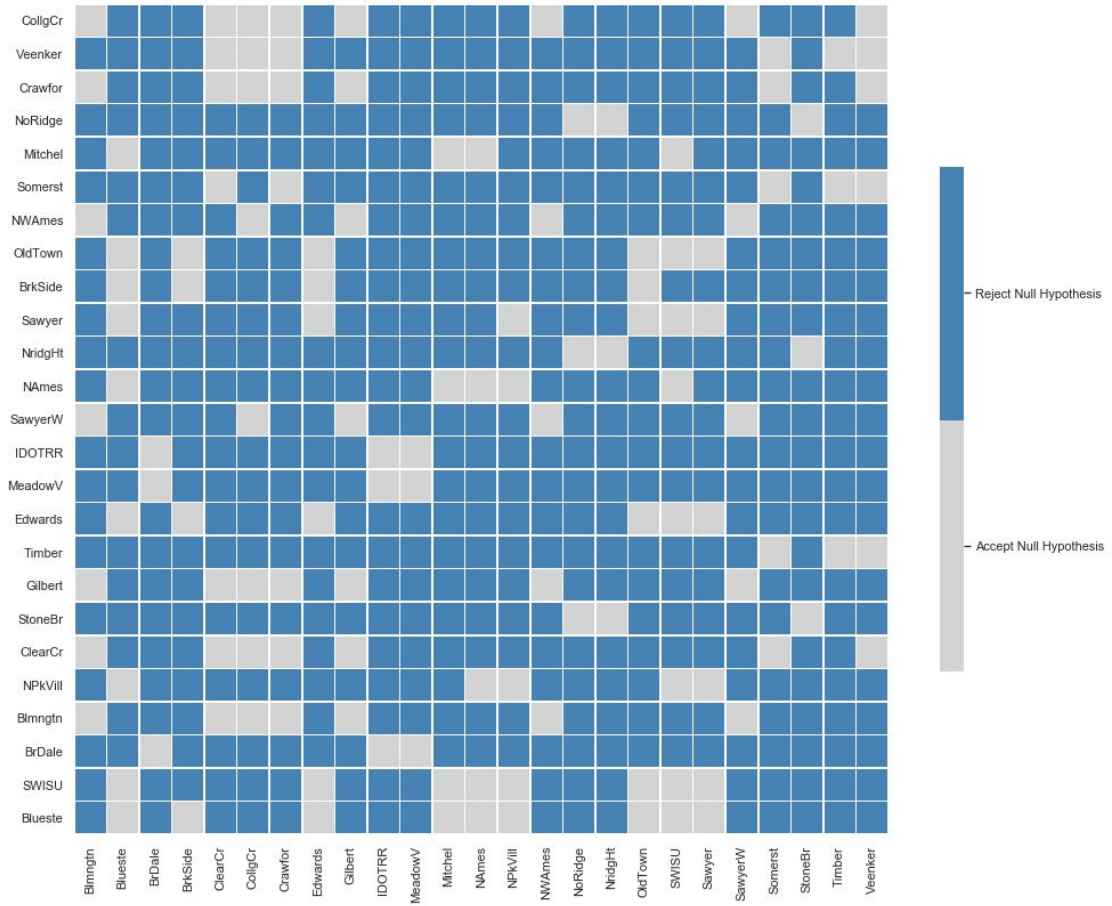


Other neighborhood Comparisons

To compare other neighborhoods, we use a python script to find all combinations of neighborhoods.

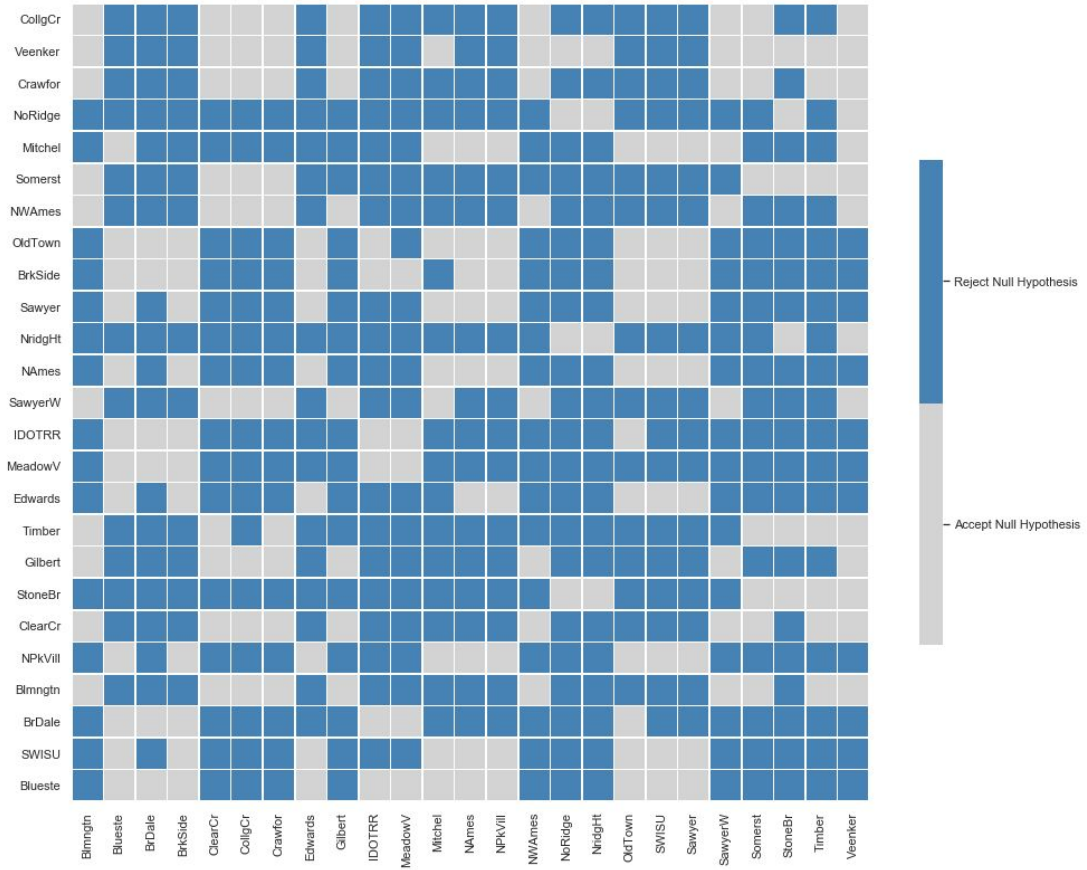
We then calculate the probability of observed mean variance using the same method applied in the previous slide.

Finally we summarize the results (for failure/acceptance of Ho) on a seaborn binary heatmap



Bonferroni Method

- Since we have so many comparisons to perform, the likelihood of committing a Type I error (rejection of a true null hypothesis or Ho) is rather high.
- To treat for this, we apply the Bonferroni correction which states to divide the significance level α by the number of comparisons.
- In this case we have .05/625 = .00008 or .008%
- We then rerun the comparison function



Conclusions

- 494 statistically significant mean differences were found before the Bonferroni method
- 414 statistically significant mean differences found after application of the method
- The value still shows a majority of the neighborhoods display significant mean differences
- **totalSF** and **OverallQual** both have high correlations to **SalePrice**
- Its yet to be proven that **totalSF** is a stronger indicator, however it does have a higher calculated Pearson coefficient
- To establish a true predictive model, some machine learning will need to be applied.