Assignment 1

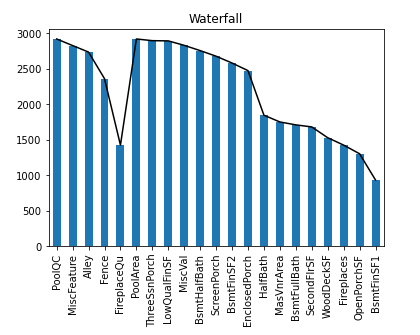
April 2021

(1) A Data Survey

The Ames Data set is quite constructed, and deals with residential housing info for Ames Iowa. The data has information about the neighborhoods and zoning for the homes, as well as size, year built and condition and quality of the homes. It is a very diverse data set but most of it revolves around features in the home like kitchen, garage, and basements. It is supposed to represent features that would affect the values for each home. The data, In constructing linear regression models, data must be cleansed and trimmed in order to build the best fitting model and create the best results. There seems to be an ample amount of data to choose from and would have to be reduced in order to create a well fit, accurate model. Most of the data appears to be observations that would be excluded from a regression model. However, there does seem to be a lot of good data about features in the home which would affect the value. We can answer a lot of questions giving the data we have, most revolving around value from the home. We can look at specific dependent variables and compare them to independent variable of sales price. You can also look at things like year built versus neighborhood or lot type versus garage size. There are numerous questions that one can answer with the data! One does need to be careful here because there is so much, it is important to do proper data cleansing by removing data columns with null or missing values or not much data in them. Its also important to find the correlation of the columns and how they fit against what the independent variable one is trying to find for.

(2) Define the Sample Population

When building a sample population it is imperative to find which columns of data are appropriate for regression analysis and which are not. To do this, one must view the quality of the data. I have done that here by viewing the columns with the most nulls and sorted them by descending order. Here I can see there are 5 columns with more than half the data are nulls. These are great candidates to be removed from the sample data. In addition, I counted the columns with the most values of zero. This means that although there are values for these columns, there is no tangible data here. I know one-hot encoding hasn’t been applied, so I know none of these are categorical values that have been converted to numerical values – they are all columns of data with a large amount of 0’s or missing data. These 16 columns will also be dropped. I add them with the previous columns for nulls and create the waterfall chart:

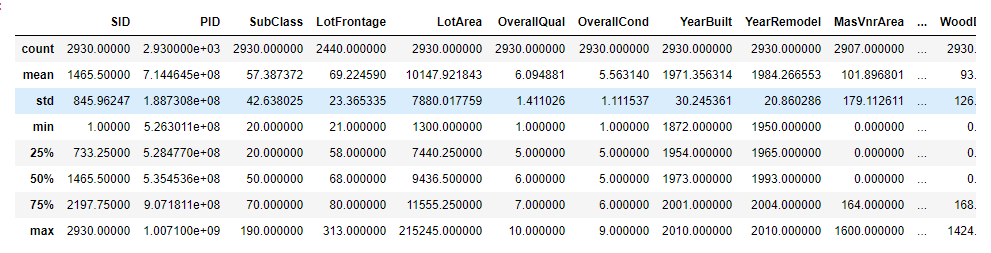


I have set the x axis as the column names, while the y axis shows the amount of nulls and zeros. The total amount of data rows for the Ames data is

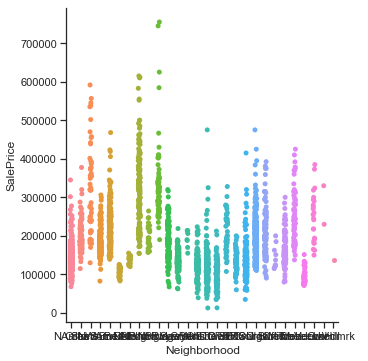
a little less than 3,000 so we can see that a lot the data in the waterfall chart is not good enough for our regression model. The waterfall chart shows the columns that meet the drop conditions, a large number of missing or incomplete data.

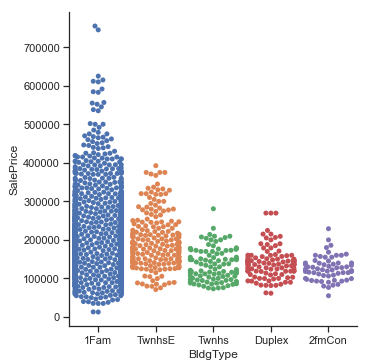
(3) A Data Quality Check

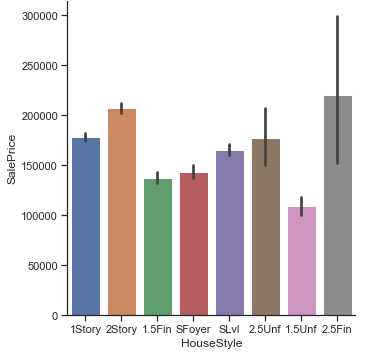
After completing the initial EDA review, I focused in on the remaining columns of data that had null or missing values. I viewed the data even further by applying the describe function to the dataset to see the top values:



The data is a mix of categorical and numerical columns, but the describe function does a good job of getting standard deviation, min and max, and different percentiles. In noting that are a lot of these columns are categorical, it is important to graph them to compare them against the main independent variable of Sales Price:





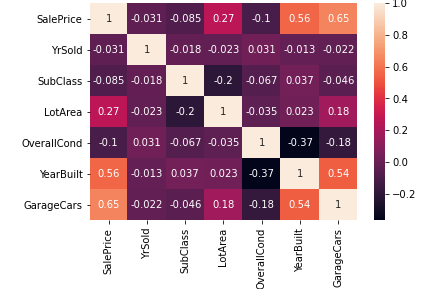


In reviewing the data to reduce the data set to 20 variables, one must be taken as the independent variable to measure value against, Sales Price. By looking at the describe model of the dataset, and some of these graphs mapping categorical values I reduced the data to values revolving specifically around the home. These variables deal with the specifics of the house itself: garage cars amount, exterior condition, paved drive, and utilities. Other values deal with data about the conditions around the house like Zoning, neighborhoods, building types, housing style and year sold. Here are the twenty columns I chose:

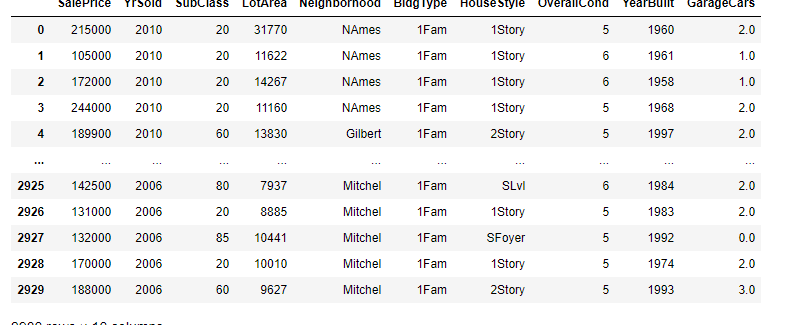


(4) An Initial Exploratory Data Analysis

For the next step, I reduce the amount of variables even further to 10. In the last part, I compared numerous categorical variables to sales price and found a few that were interesting. I also did a deeper dive into the data with the describe function. In order to see how these fit best, I created a correlation matrix to see the columns that are highest correlated against one each other which show the data which affects the sales price the most:



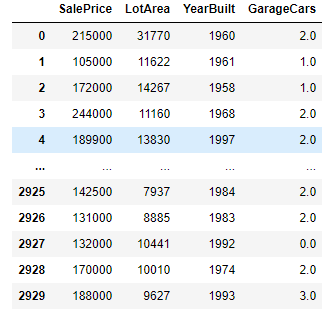
These values do not apply to categorical values because the correlation matrix only compares numerical values, but it shows a good job of showing how the top numerical values are correlated to sales price. To reduce the top 20 to the top 10, I chose the highest correlated values from the previous categorical data review, and the top values from the correlation matrix to make this top 10:



The main discrete variables are garage cars, year built and lot area, while the main categorical values are Neighborhood and Building Type.

(5) An Initial Exploratory Data Analysis for Modeling

The final task, which proves to be the most difficult task, is to reduce the final ten variables down to three. This is difficult because after conducting an EDA there are so many great variables to choose from, and some of them are very close in terms of impact upon sales Price. In viewing the categorical variables, I eliminated the variable Building type because the highest values homes were all in the single family home category. The other categories were distinctly less valuable than this first category and thus would seem clear to have no impact. Next, I eliminated House Style because the data was too similar amongst all the categories. There was not a single category that stood out amongst the rest, instead values seemed to be evenly disparate despite house style. The last categorical value, neighborhood, has values that show different neighborhoods having an affect on sales price, and thus should be kept. The discrete values can easily be picked by looking at the correlation matrix and selecting the top values. Here, the top values are Garage Cars, Year built and Lot area having the highest correlation. Garage cars and year built have very high correlation both over .5. The response variable is Sales Price. Everything that is value correlated can be tied to sales price, and it is the best marker of value. The EDA does not show potential concerns, in fact the opposite, that there are numerous values that have high correlation and can be used to create an effective model. Currently the EDA does not show there needs to be a transformation of the sales price as there are variables with high correlation. However, if one wanted to use another variable as value transformation could be sued to create high correlated variables, or even tried to get sales price more correlated. One does run the risk of overfitting if they do that though.





(6) Summary/Conclusions:

After completing the EDA and narrowing down the variables to three, the final variables I chose were all discrete: Garage Cars, Lot Area, and Year built. I chose not to use the categorical variable even though it had an affect on the sales price because the correlation seemed higher with the lot area. Year Build and garage Cars both had correlation above .5 so they had to be included because they suggest a high correlation to sales price. I don’t see any potential problems as the data variables have high correlation so we should get a good result. We also have not transformed the response variable so this eliminates the potential for overfitting. If we had changed parts of the sales price to mean or another value for sales price, this could have left data which was more skewed to our sample dataset. Due to the high correlation and discrete variable nature of the dataset, there would not be a further need to transform the data.