Assignment 1

April 2021

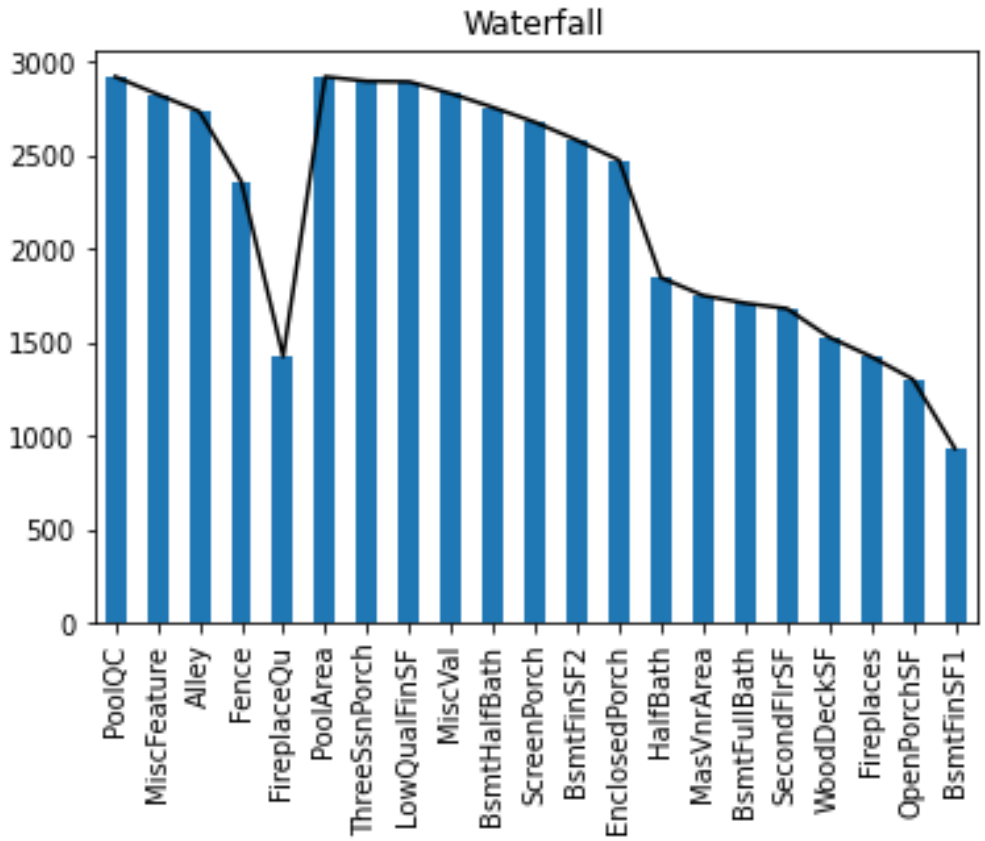
(1) A Data Survey

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

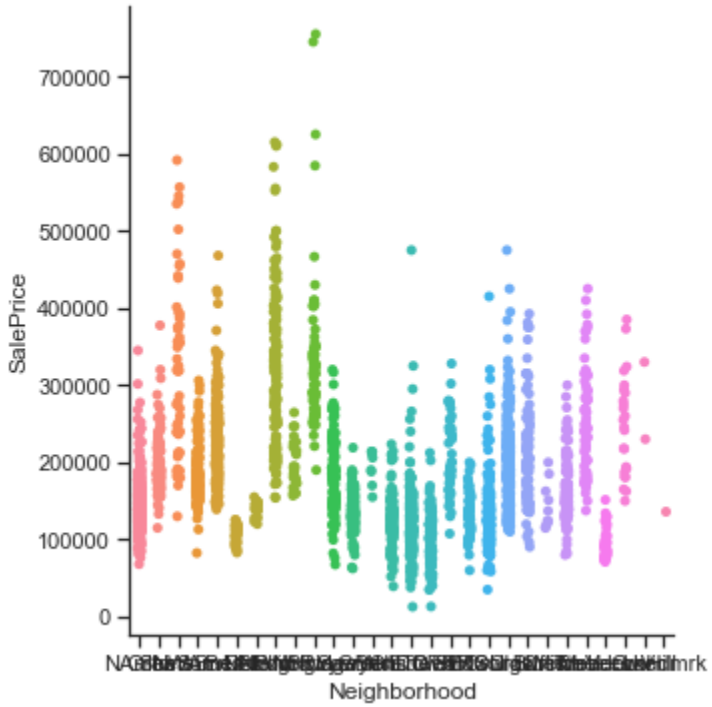
a little less than 3,000 so we can see that a lot the data in the waterfall chart is not good enough for our regression model. The waterfall chart shows the columns that meet the drop conditions, a large number of missing or incomplete data.

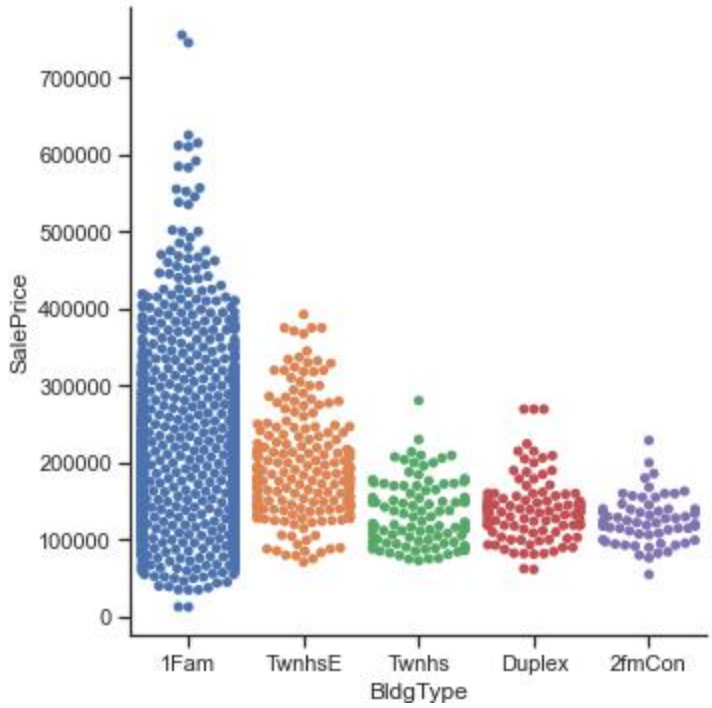
(3) A Data Quality Check

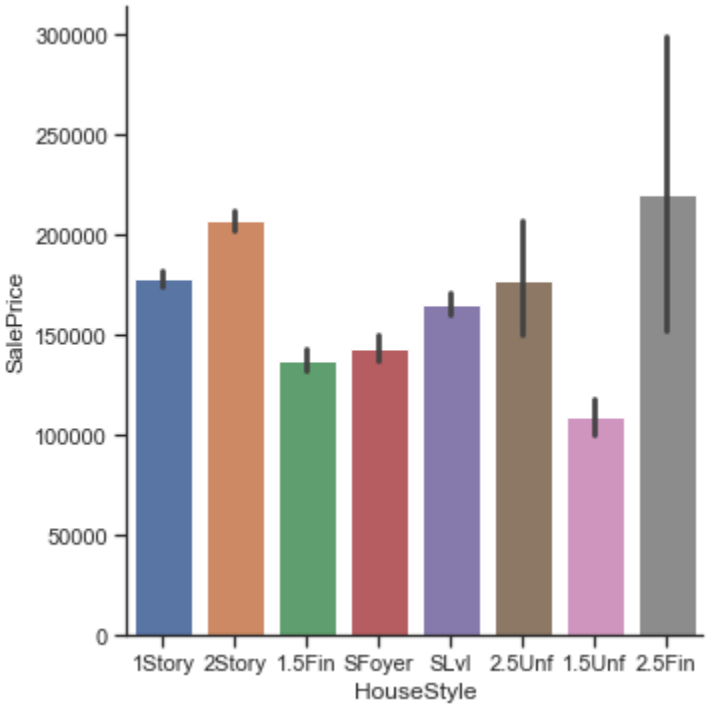
After completing the initial EDA review, I focused in on the remaining columns of data that had null or missing values. I viewed the data even further by applying the describe function to the dataset to see the top values:

	SID	PID	SubClass	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	YearRemodel	MasVnrArea	...	WoodI
count	2930.00000	2.930000e+03	2930.000000	2440.000000	2930.000000	2930.000000	2930.000000	2930.000000	2930.000000	2907.000000	...	2930.
mean	1465.50000	7.144645e+08	57.387372	69.224590	10147.921843	6.094881	5.563140	1971.356314	1984.266553	101.896801	...	93.
std	845.96247	1.887308e+08	42.638025	23.365335	7880.017759	1.411026	1.111537	30.245361	20.860286	179.112611	...	126.
min	1.00000	5.263011e+08	20.000000	21.000000	1300.000000	1.000000	1.000000	1872.000000	1950.000000	0.000000	...	0.
25%	733.25000	5.284770e+08	20.000000	58.000000	7440.250000	5.000000	5.000000	1954.000000	1965.000000	0.000000	...	0.
50%	1465.50000	5.354536e+08	50.000000	68.000000	9436.500000	6.000000	5.000000	1973.000000	1993.000000	0.000000	...	0.
75%	2197.75000	9.071811e+08	70.000000	80.000000	11555.250000	7.000000	6.000000	2001.000000	2004.000000	164.000000	...	168.
max	2930.00000	1.007100e+09	190.000000	313.000000	215245.000000	10.000000	9.000000	2010.000000	2010.000000	1600.000000	...	1424.

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

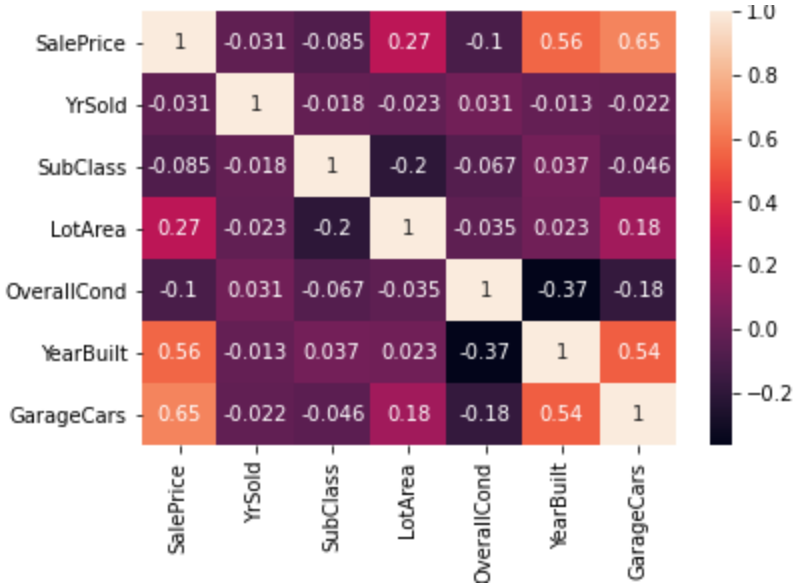
of these graphs mapping categorical values I reduced the data to values revolving specifically around the home. These variables deal with the specifics of the house itself: garage cars amount, exterior condition, paved drive, and utilities. Other values deal with data about the conditions around the house like Zoning, neighborhoods, building types, housing style and year sold. Here are the twenty columns I chose:

	SalePrice	YrSold	GarageCond	SubClass	Zoning	LotArea	Utilities	Neighborhood	Condition1	BldgType	HouseStyle	OverallCond	YearBuilt	Ext
0	215000	2010	TA	20	RL	31770	AllPub	NAmes	Norm	1Fam	1Story	5	1960	
1	105000	2010	TA	20	RH	11622	AllPub	NAmes	Feedr	1Fam	1Story	6	1961	
2	172000	2010	TA	20	RL	14267	AllPub	NAmes	Norm	1Fam	1Story	6	1958	
3	244000	2010	TA	20	RL	11160	AllPub	NAmes	Norm	1Fam	1Story	5	1968	
4	189900	2010	TA	60	RL	13830	AllPub	Gilbert	Norm	1Fam	2Story	5	1997	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
2925	142500	2006	TA	80	RL	7937	AllPub	Mitchel	Norm	1Fam	SLvl	6	1984	
2926	131000	2006	TA	20	RL	8885	AllPub	Mitchel	Norm	1Fam	1Story	5	1983	
2927	132000	2006	NaN	85	RL	10441	AllPub	Mitchel	Norm	1Fam	SFoyer	5	1992	
2928	170000	2006	TA	20	RL	10010	AllPub	Mitchel	Norm	1Fam	1Story	5	1974	
2929	188000	2006	TA	60	RL	9627	AllPub	Mitchel	Norm	1Fam	2Story	5	1993	

2930 rows × 19 columns

(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:

	SalePrice	YrSold	SubClass	LotArea	Neighborhood	BldgType	HouseStyle	OverallCond	YearBuilt	GarageCars
0	215000	2010	20	31770	NAmes	1Fam	1Story	5	1960	2.0
1	105000	2010	20	11622	NAmes	1Fam	1Story	6	1961	1.0
2	172000	2010	20	14267	NAmes	1Fam	1Story	6	1958	1.0
3	244000	2010	20	11160	NAmes	1Fam	1Story	5	1968	2.0
4	189900	2010	60	13830	Gilbert	1Fam	2Story	5	1997	2.0
...	...	...	...	...	...	...	...	...	...	...
2925	142500	2006	80	7937	Mitchel	1Fam	SLvl	6	1984	2.0
2926	131000	2006	20	8885	Mitchel	1Fam	1Story	5	1983	2.0
2927	132000	2006	85	10441	Mitchel	1Fam	SFoyer	5	1992	0.0
2928	170000	2006	20	10010	Mitchel	1Fam	1Story	5	1974	2.0
2929	188000	2006	60	9627	Mitchel	1Fam	2Story	5	1993	3.0

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.

	SalePrice	LotArea	YearBuilt	GarageCars
0	215000	31770	1960	2.0
1	105000	11622	1961	1.0
2	172000	14267	1958	1.0
3	244000	11160	1968	2.0
4	189900	13830	1997	2.0
...	...	...	...	...
2925	142500	7937	1984	2.0
2926	131000	8885	1983	2.0
2927	132000	10441	1992	0.0
2928	170000	10010	1974	2.0
2929	188000	9627	1993	3.0

	SalePrice	LotArea	YearBuilt	GarageCars
count	2930.000000	2930.000000	2930.000000	2929.000000
mean	180796.060068	10147.921843	1971.356314	1.766815
std	79886.692357	7880.017759	30.245361	0.760566
min	12789.000000	1300.000000	1872.000000	0.000000
25%	129500.000000	7440.250000	1954.000000	1.000000
50%	160000.000000	9436.500000	1973.000000	2.000000
75%	213500.000000	11555.250000	2001.000000	2.000000
max	755000.000000	215245.000000	2010.000000	5.000000

(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.