Feature Engineering Project Write up

This project aimed to determine which factors were most important to home buyers when purchasing a home. Both numerical and categorical factors were considered. Multiple data visualization methods, including pair plots, count plots, and heat maps, were implemented to determine the most significant factors.

Numerical Data

Initial list of significant data made using guesses

Let us begin by arranging the data columns in order of decreasing significance in house purchases. This list is made using my judgment and opinion, not with previous analysis. The most important factors are given a rating of 10, and the least important are given the rating of 1.

Most important columns

SalePrice (The property's sale price is in dollars. **This is the target variable that we're trying to predict**)

Utilities (Type of utilities available). 10
OverallCond (Overall condition rating). 10
Functional (Home functionality rating). 10
SaleCondition (Condition of sale). 10
Electrical (Electrical system). 10
Bedroom (Number of bedrooms above basement level). 10
FullBath (Full bathroom above grade). 10
Neighborhood (Physical locations within Ames city limits). 10

People care a great deal about the number of bedrooms and bathrooms. They also want lots of utilities, like laundry machines, dryers, an air conditioner, electricity, built-in lights, water, etc. Obviously, they want high quality homes that look good and have reliable electrical and utility systems. When considering buying a house, people want the homes that look the best and feel the most comfortable to live in.

Also, living in a good neighborhood is important since that determines where you can work, go the school, and go out for fun. Property values depend a great deal on location. From the perspective of the home buyer, neighborhoods cannot be ignored since they don't stay indoors all the time.

OverallQual (Overall material and finish quality). 9
CentralAir (Central air conditioning). 9
1stFlrSF (First Floor square Feet). 9
2ndFlrSF (Second floor square feet). 9

High quality material and air conditioning for hot days is very important to many, especially those living in hot climates. Naturally, people also care about the size of the home, but what matters more is the number of rooms.

YearBuilt (Original construction date). 8

ExterQual (Exterior material quality). 8
ExterCond (The present condition of the material on the exterior). 8

The year the home was built can determine how old the utilities are. If it’s too old, the utilities might break down and need to be fixed. That takes time, money, and headaches.

The quality of appearance is mostly about aesthetics. Important, but not as practical.

Important, but less so

KitchenQual (Kitchen quality). 7

The kitchen should have the bare minimum number of utilities. Large kitchens are good since people want to make their own meals. But people don’t care as much about this.

GarageType (Garage location). 6

The garage should be close enough that you can walk to it in five minutes or less. This doesn’t matter too much unless you’re carry heavy or delicate objects from the home to the car or vice versa.

Not very significant

BsmtCond (Condition of basement). 4

The basement just needs to look good. But it’s mostly for storage.

HalfBath (Half bathrooms above grade). 3
BsmtQual (Height of the basement). 3

Half bathroom are optional. The basement doesn’t need to be too high. Just high enough that there’s enough space for you to stand up straight and not have to worry about bumping your head on the ceiling.

Kitchen (Number of kitchens). 2
Heating (Type of heating). 2

There only needs to be one kitchen and people can buy their own heaters.

All other columns are less significant than these ones and can therefore be ignored for the purposes for analysis.

Most useful numerical columns based on graph skewness

Left skewed (tail extends left):

YearBuilt
GarageYrBlt

Right skewed (tail extends right):

LotFrontage
LotArea
BsmtFinSF1

BsmtUnfSF
TotalBsmtSF
1stFlrSF
GrLivArea
TotRmsAbvGrd
GarageArea
WoodDeckSF
OpenPorchSF
MoSold
SalePrice

Numerical Correlation Matrix

The correlation values for SalePrice are located in the top row and the leftmost column. Having high correlation values in that row/column is good. However, we may encounter some instances of two variables being highly correlated with each other rather than the SalePrice. This is undesirable since that means we have variables that are redundant for our analysis. If two variables are highly correlated with each other, any analysis performed on one will yield very similar results to an analysis performed on the other.

We must identify and remove redundant variables. If we define two variables that have a correlation of at least 0.8 as being redundant, then we can say the following variables are redundant.

TotRmsAbvGrd and GrLivArea have 0.83 correlation.

TotRmsAbvGrd is the total number of rooms above ground that are not bathrooms.
GrLivArea is the total above ground living area in square feet.

This correlation is expected. After all, the more above ground living space you have, the more rooms you can put above the ground.

TotRmsAbvGrd has 0.53 correlation with SalePrice and GrLivArea has 0.71 correlation with SalePrice. We will eliminate TotRmsAbvGrd since it's correlation with SalePrice is lower than GrLivArea.

GarageArea and GarageCars have 0.88 correlation.

GarageArea is the size of the garage in square feet
GarageCars is the size of the garage in car capacity

Both variables are concerned with the size of the garage. The way the garage size is measured is different for each variable, but they are both measuring the same part of the home.

GarageArea has 0.62 correlation with SalePrice and GarageCars has 0.64 correlation with SalePrice. Thus, we will eliminate GarageArea, which has the lower correlation.

1stFlrSF and TotalBsmtSF have 0.82 correlation.

1StFlrSF is the size of the first floor in square feet.
TotalBsmtSF is the area of the basement in square feet.

The basement is located right underneath the first floor. Naturally, the size of the basement is largely constrained by how large the first floor is allowed to be. There is some leeway to make the basement bigger than the first floor since it is located underground, but the first floor's size is significant confinement metric.

1StFlrSF has 0.61 correlation with SalePrice and TotalBsmtSF also has 0.61 correlation with SalePrice. We can eliminate either one, so let's eliminate TotalBsmtSF since every home has a first floor but not every home has a basement.

Our final set of correlated variables that is nor redundant is

SalePrice, OverallQual, GrLivArea, GarageCars, 1stFlrSF, FullBath, YearBuilt

Pairplot of Numerical Data

Use the final set of important columns in a pairplot to view all the correlations in one picture. The final column set is:

SalePrice, OverallQual, GrLivArea, GarageCars, 1stFlrSF, FullBath, YearBuilt

The middle diagonal of plots from the top left graph to the bottom right graph are just plots of the same variables matched with each other. We can ignore these graphs.

To check if there is a high correlation between variables, check how many plots have data points that roughly follow a positively-sloped best fit line. The more points close to that line, the stronger the correlation between the variables. The correlation will be strongest for best fit lines that have a 45 degree angle with x-axis.

Here is how all the variables compare with SalePrice on the y-axis

OverallQual: Most data points are located on the 45 degree best fit line or below it. This is sensible since most home will be valued proportional to their exterior quality. It would be possible to get the home for a cheaper value due to other circumstances, but few home buyers would be willing to pay for a home that has a higher price that its quality. As such, few data points are located above the 45 degree best fit line.

Note that OverallQual is concerned with the aesthetics of the home. Does the home look good? People are very concerned with appearances, which is why the correlation is so strong.

GrLivArea: As with OverallQual, points are located to the right or below of the best fit line, which has a positive slope. The best fit line looks to have a 60 degree angle.

More living space means more land in the home. Naturally, people on the lookout for bigger homes would be willing to pay more money.

GarageCars: Since garages have a small, discrete number of cars, the shape of this graph is similar to a histogram. It makes sense to see that homes with high-capacity garages will be worth more since buyers would be paying for more land. The best fit line looks to have a 45 degree angle.

1stFlrSF: The best fit line looks to be 60 degrees. Most points are on the line, to the right of it, or below it. This is understandable for reasons similar to the above variables. More living space means a higher cost. Also, some homes have only one floor without an attic or basement, which makes the first floor the most important floor to consider.

FullBath: Since there are only a small number of bathrooms in a home, the shape of this graph is similar to a histogram. It makes sense to see that homes with more bathrooms will be worth more since buyers would be paying for more land. Also, more bathrooms is a huge convenience in cases plumbing work needs to be done on the facilities in one restroom. The best fit line looks to have a 45 degree angle.

YearBuilt: There is a best fit line with a 30 degree angle with the x-axis for all homes except very recently built ones. The reason the best-fit line is 30 degrees for most homes, as opposed to 45 degrees or more, is that people care about a home's age so much that if it is too old, then few will pay high prices for it. Every home has an age, and unlike the other variables, the age of a home increases with every year. Older homes have older facilities and utilities, increasing the chance that maintenance will be needed to fix something. No matter how high quality or spacious a home is, the older it is, the more wary people are when considering a purchase.

Note that for the right side of the graph, where the homes are recent, the data points are much higher on the x-axis, indicating a higher price for such homes. This is sensible since very recently built homes are the least likely to need maintenance. Also, these new homes will have the latest models of utilities like dishwashers, bathrooms, ranges, microwave ovens, laundry machines, and dryers. The prestige and efficiency of such technologies is attractive enough that people will seek out homes that possess such facilities.

Categorical Data

Missing value treatment

The following columns have missing values:

Alley, MasVnrType, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, Electrical, FireplaceQu, GarageType, GarageFinish, GarageQual, GarageCond, PoolQC, Fence, MiscFeature

Alley: Type of alley access
MasVnrType: Masonry veneer type
BsmtQual: Height of the basement
BsmtCond: General condition of the basement
BsmtExposure: Walkout or garden level basement walls
BsmtFinType1: Quality of the basement finished area
BsmtFinType2: Quality of second finished area (if present)
Electrical: Electrical system

FireplaceQu: Fireplace quality
GarageType: Garage location
GarageFinish: Interior finish of the garage
GarageQual: Garage quality
GarageCond: Garage condition
PoolQC: Pool quality
Fence: Fence quality
MiscFeature: Miscellaneous feature not covered in other categories

Note that for many of the columns with missing values, the lack of data points may be indicative of lack of features. These missing-value columns are all mostly concerned with features that not all homes have. Not every home has a basement, fireplace, garage, fence, pool, alley access, or miscellaneous features. In other words, the missing values are just placeholders for the term “not applicable”.

Every home will have an electrical system, but the type of electrical system doesn’t matter much to home buyers so much as the presence of a functional one. The category of electrical system may not be important.

The type of masonry in a home may not specified. But every home is made of some kind of material. The missing data points indicates that the type of masonry veneer is not specified.

We can conclude that only the missing values in the Electrical and MasVnrType columns are significant and the other missing values can be explained by lack of features.

Categorical data Countplots

In many of the countplots, only one or a small number of possible values has a large count. The counts are not evenly distributed across the different values. This indicates that many of the homes have very similar categorical properties. We can infer from these graphs that home buyers are all looking to purchase homes with similar features, so the categorical data may not be useful in distinguishing what is appealing aside from identifying the most popular categorical properties.

p-value and Chi-squared test

The values for the p-value and Chi-square test are both “nan”, so the values are not numbers. This result indicates homogeneity in the data and we should ignore both the p-value and chi-squared tests for our analysis.

Significant Categorical and Numerical Values

From the Countplots, the only values that are significant are the ones with a large spread of data instead of high concentrations into one bin. If one bin has too high of a concentration, extracting useful information from it will not give anything useful since the data will be almost the same.

LotConfig, Neighborhood, HouseStyle, RoofStyle, Exterior1st, Exterior2nd, MasVnrType, ExterQual, Foundation, BsmtQual, BsmtFinType1, HeatingQC, KitchenQual, FireplaceQu, GarageType, GarageFinish, PoolQC, Fence

Final Results

Those are most important categorical data columns. From above, our most important numerical data columns, according to the pair plot, are

SalePrice, OverallQual, GrLivArea, GarageCars, 1stFlrSF, FullBath, YearBuilt

The most significant categorical data columns are

LotConfig, Neighborhood, HouseStyle, RoofStyle, Exterior1st, Exterior2nd, MasVnrType, ExterQual, Foundation, BsmtQual, BsmtFinType1, HeatingQC, KitchenQual, FireplaceQu, GarageType, GarageFinish, PoolQC, Fence

We can combine these columns into one DataFrame for analysis with a machine learning model. But we need to perform some kind of encoding, such as label encoding or one-hot encoding, to use the categorical data in a machine learning model. If we encode the categorical data, then our final DataFrame will be ready for processing.