**House Price Prediction Project Report**

1. Business Understanding
   1. Objective

Predict the house price

* 1. Description

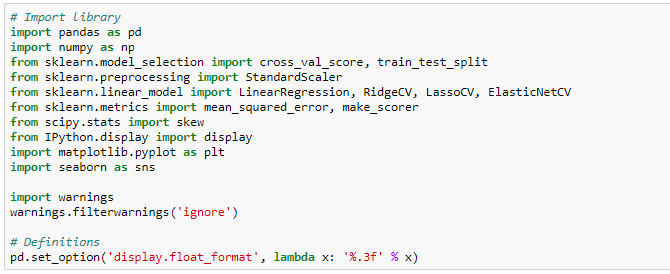
Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

In this project, I will complete the analysis of which house feature will impact the house price. In particular, I will apply machine learning to predict the house price.

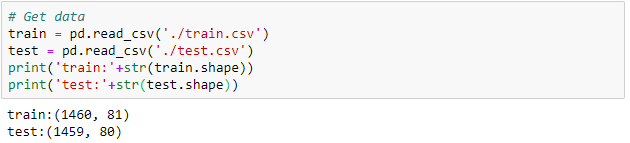
1. Data Understanding
   1. Import library

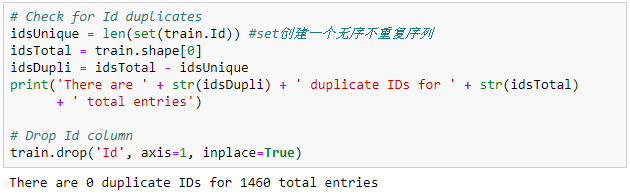
First of all, I need import python libraries containing the necessary functionality will need.



* 1. Load data

Next, load train.csv and test.csv dataset. Have a peek at the dataset size and variables.







* 1. Statistical summaries and visualization

To understand the data, I am going to consider some key facts about various variables including their relationship with the target variables, i.e. saleprice.

Here is the variable descriptions:

SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.

MSSubClass: The building class

MSZoning: The general zoning classification

LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

Street: Type of road access

Alley: Type of alley access

LotShape: General shape of property

LandContour: Flatness of the property

Utilities: Type of utilities available

LotConfig: Lot configuration

LandSlope: Slope of property

Neighborhood: Physical locations within Ames city limits

Condition1: Proximity to main road or railroad

Condition2: Proximity to main road or railroad (if a second is present)

BldgType: Type of dwelling

HouseStyle: Style of dwelling

OverallQual: Overall material and finish quality

OverallCond: Overall condition rating

YearBuilt: Original construction date

YearRemodAdd: Remodel date

RoofStyle: Type of roof

RoofMatl: Roof material

Exterior1st: Exterior covering on house

Exterior2nd: Exterior covering on house (if more than one material)

MasVnrType: Masonry veneer type

MasVnrArea: Masonry veneer area in square feet

ExterQual: Exterior material quality

ExterCond: Present condition of the material on the exterior

Foundation: Type of foundation

BsmtQual: Height of the basement

BsmtCond: General condition of the basement

BsmtExposure: Walkout or garden level basement walls

BsmtFinType1: Quality of basement finished area

BsmtFinSF1: Type 1 finished square feet

BsmtFinType2: Quality of second finished area (if present)

BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area

TotalBsmtSF: Total square feet of basement area

Heating: Type of heating

HeatingQC: Heating quality and condition

CentralAir: Central air conditioning

Electrical: Electrical system

1stFlrSF: First Floor square feet

2ndFlrSF: Second floor square feet

LowQualFinSF: Low quality finished square feet (all floors)

GrLivArea: Above grade (ground) living area square feet

BsmtFullBath: Basement full bathrooms

BsmtHalfBath: Basement half bathrooms

FullBath: Full bathrooms above grade

HalfBath: Half baths above grade

Bedroom: Number of bedrooms above basement level

Kitchen: Number of kitchens

KitchenQual: Kitchen quality

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

Functional: Home functionality rating

Fireplaces: Number of fireplaces

FireplaceQu: Fireplace quality

GarageType: Garage location

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

GarageQual: Garage quality

GarageCond: Garage condition

PavedDrive: Paved driveway

WoodDeckSF: Wood deck area in square feet

OpenPorchSF: Open porch area in square feet

EnclosedPorch: Enclosed porch area in square feet

3SsnPorch: Three season porch area in square feet

ScreenPorch: Screen porch area in square feet

PoolArea: Pool area in square feet

PoolQC: Pool quality

Fence: Fence quality

MiscFeature: Miscellaneous feature not covered in other categories

MiscVal: $Value of miscellaneous feature

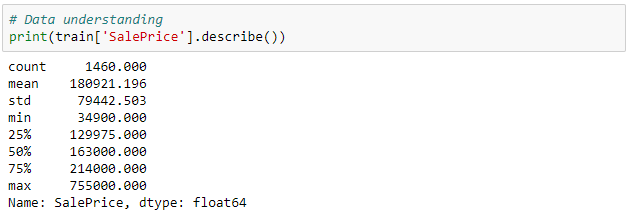
MoSold: Month Sold

YrSold: Year Sold

SaleType: Type of sale

SaleCondition: Condition of sale

We can see from below, the saleprice is range from 34900 to 755000, while 75% of the price is lower than 214000. Also, the skewness is bigger than 0, which means most of the saleprice located in low price.





1. Preprocessing
   1. Investigating numerical variables

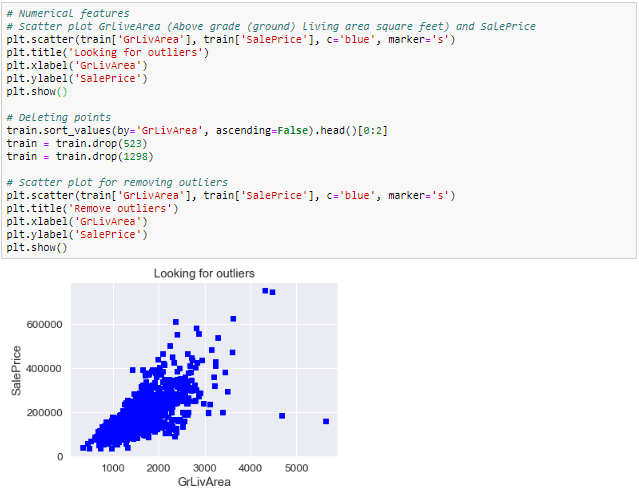
There are 2 kinds of variables, one is numerical variable and the other is categorical variable. Numerical variable is one with values of integers or real numbers, while a categorical variable is a variable that can take on one of a limited , and usually fixed number of possible values.

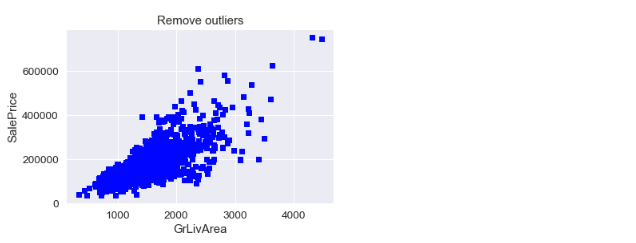
Since there are 80 variables in this case, I want to select several variables to see their relationship with saleprice.

First, I start from numerical variables. Here I selected ‘GrLivArea’ , it seems to be linearly related with ‘SalePrice’, it’s positive, which means that as one variable increases, the other also increases.

What has been revealed from below scatter plot:

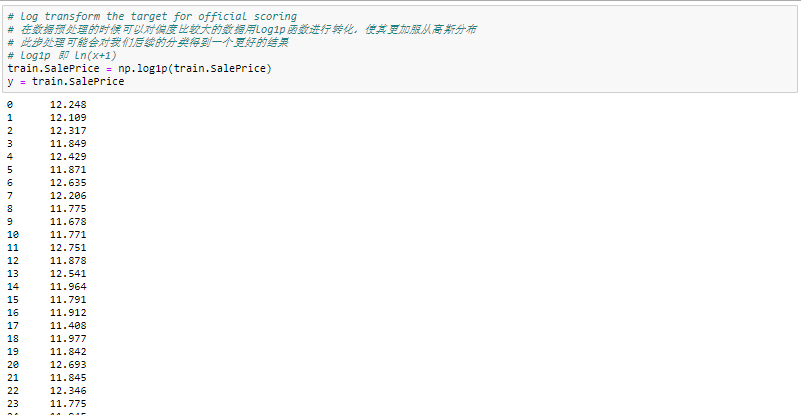
* The two values with bigger ‘GrLivArea’ seem strange and they are not following the crowd. We can speculate why this is happening. Maybe they refer to agricultural area and that could explain the low price. I’m not sure about this but I’m quite confident that these two points are not representative of the typical case. Therefore, we’ll define them as outliers and delete them.
* For the two observations in the top of the plot, look like they are also two special cases. However, they seem to be following the trend. For that reason, we will keep them.





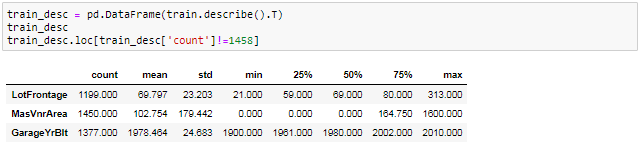
* 1. Data smoothing

Taking logs means that errors in prediction expensive houses and cheap houses will affect the result equally.



* 1. Dealing with missing data
     1. Check numerical missing data

Below code can only check numerical data, for non-numerical missing data. For categorical missing data, pls see 3.3.2



* + 1. Check both numerical and non-numerical missing data, handle missing values for features where median/mean or most common value does not make sense.



* 1. Transfer numerical features to categories features

From the dataset, we can see that some numerical features are actually really categories, so this step, let’s transfer these numerical features to categories.



* 1. Encode categories features to ordered numbers

Encode some categories features to ordered numbers when there is information in the order.



* 1. Create new features

Here I will create new features in 3 ways:

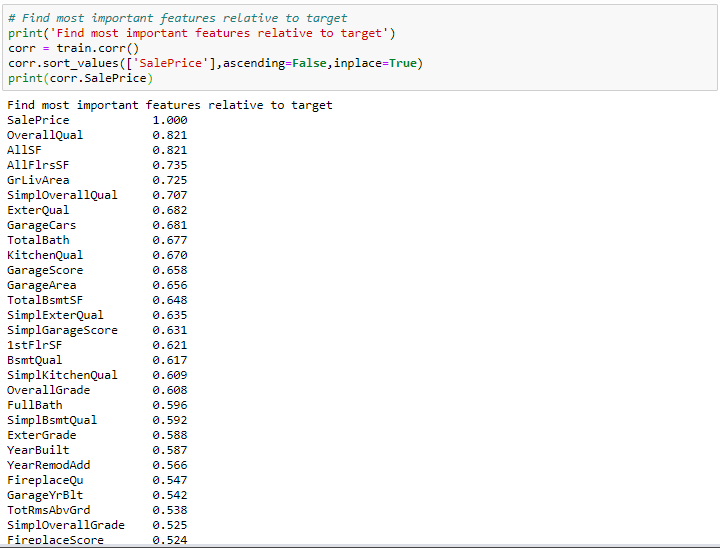
1. Simplifications of existing features
2. Combinations of existing features
3. Polynomials on the top 10 existing features

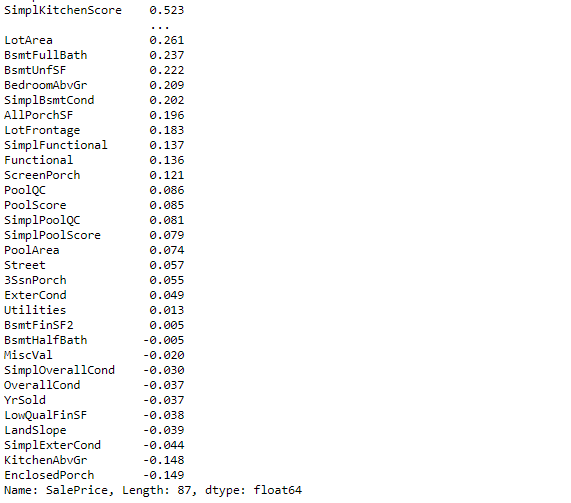




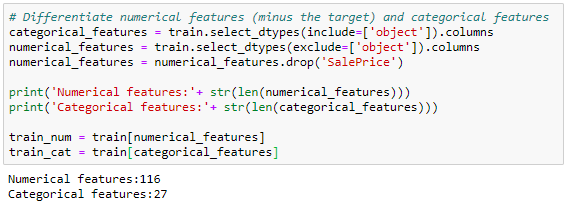


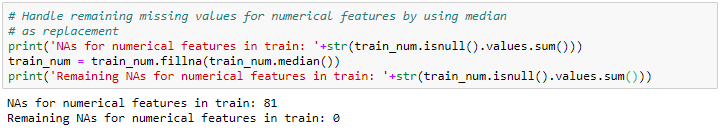






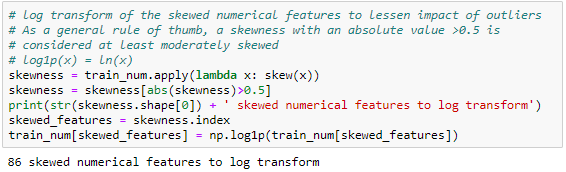




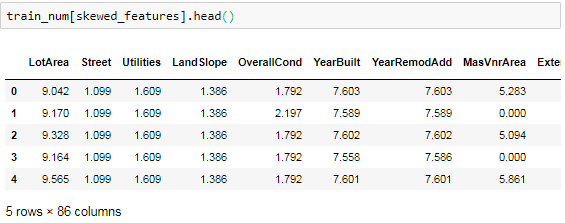


* 1. Smoothing numerical features

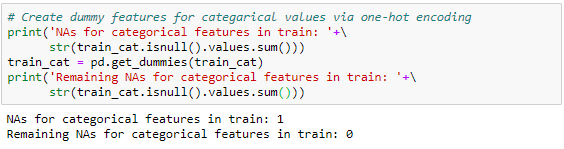
Here I use log transfer for high skewed numerical features, which can lessen impact of outliers. A skewness with an absolute value greater than 0.5, is considered at least moderately skewed.



Below is the data frame after skewed.

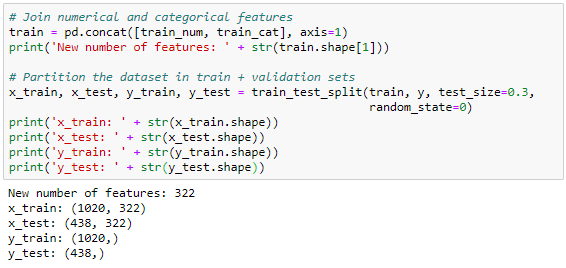


* 1. Create dummy features for categorical values



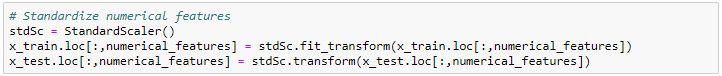
1. Modeling
   1. Preparation

Before doing modeling, join numerical and categorical features first and split test and train data.



* 1. Standardize numerical features

Note: standardize cannot be done before the partitioning, as we don’t want to fit the StandardScaler on some observations that will later be used in the test set.

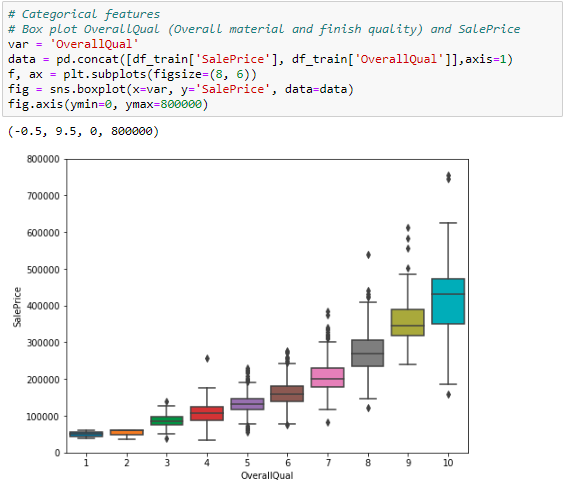


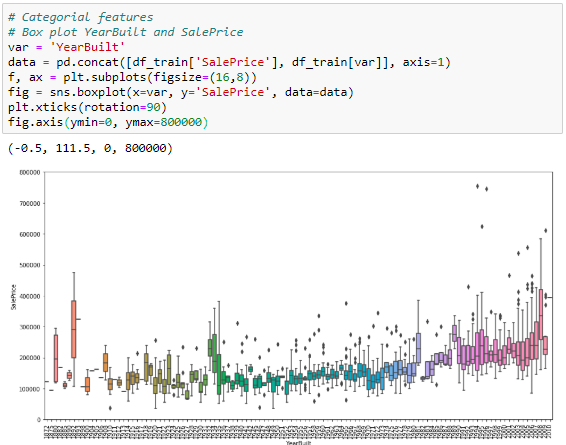
* 1. Linear Regression
     1. Linear Regression with Ridge regularization (L2 penalty)

Regularization is a very useful method to handle collinearity, filter out noise from data, and eventually prevent over fitting. The concept behind regularization is to introduce additional information (bias) to penalize extreme parameter weights.

Ridge regression is an L2 penalized model where we simply add the squared sum of the weights to our cost function.

Next step, I select categorical variables. Here I selected ‘OverallQual’ and ‘YearBuilt’, both of them also seem to be related with ‘SalePrice’. The relationship seems to be strong in the case of ‘OverallQual’, where the box plot shows how sales prices increase with the overall quality.





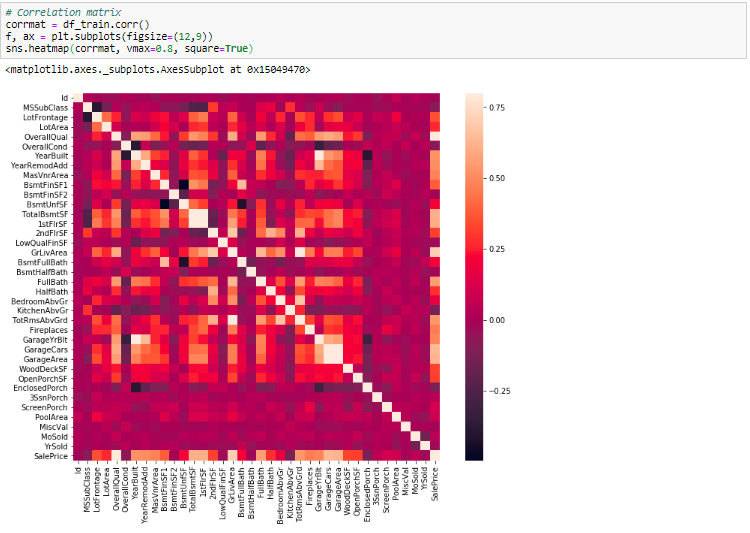
2.3.2 Variable correlation

Until now, I just follow my intuition and analyzed the variables I thought were important. But this method was subjective. As an engineer, I need a method able to withstand the winds of subjectivity. So, a heat map of correlation may give us an understanding of which variables are important, and the relationship between each variables.

To explore the universe, below is three things I plan to do:

* Correlation matrix (heatmap style)
* ‘SalePrice’ correlation matrix (zoomed heatmap style)
* Scatter plots between the most correlated variables (move like Jagger style)

2.3.2.1 Correlation matrix (heatmap style)

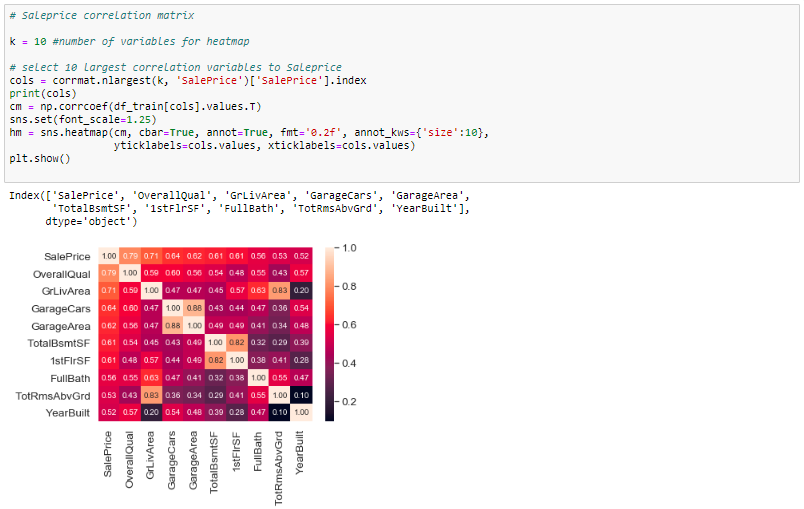


This heatmap is the best way to get a quick overview of the all variables and their relationship.

At first sight, there are 2 square that get my attention. The first one refers to the ‘TotalBsmtSF’ and ‘1stFlrSF’, and the second one refers to the ‘GarageCars’ and ‘GarageArea’ variables. Both cases show how significant the correlation is between these variables. Actually, this correlation is so strong that it can indicate a situation of multicollinearity. If we think about these variables, we can conclude that they give almost the same information so multicollinearity really occurs. Heatmap is great to detect this kind of situations and in problems dominated by feature selection.

2.3.2.2 ‘SalePrice’ correlation matrix (zoomed heatmap style)

Another thing that got my attention is the ‘SalePrice’ correlations. We can see ‘GrLivArea’, ‘TotalBsmtSF’ and ‘OverallQual’ and some other variables show strong correlations with ‘SalePrice’. That’s what I plan to do here: try to find the top 10 variables which has the strongest relationship with ‘SalePrice’.

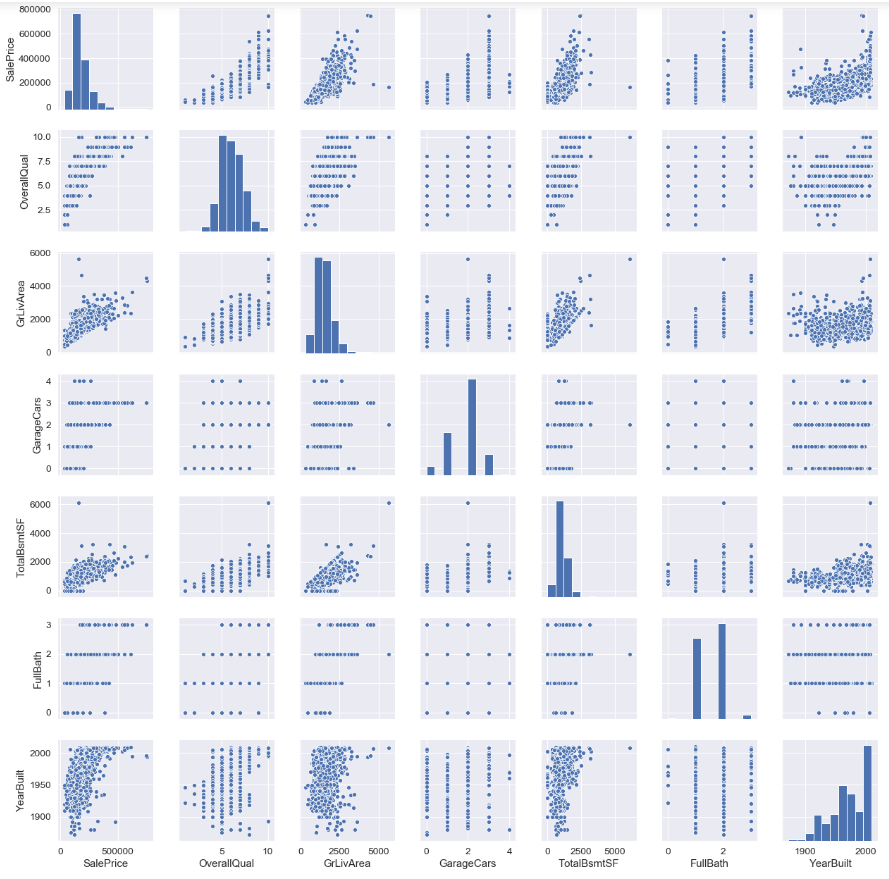


From above correlation heatmap, I find the top 10 variables most correlated with ‘SalePrice’.

* ‘OverallQual’, ‘GrLivArea’, ‘TotalBsmtSF’ are strongly correlated with ‘SalePrice’, this already be demonstrated in 2.3.1 with on the plot.
* ‘GarageCars’ and ‘GarageArea’ are also some of the most strongly correlated variables. However, the number of cars that fit into the garage is a consequence of the garage area. ‘GarageCars’ and ‘GarageArea’ are like twin brothers. You’ll never be able to distinguish them. Therefore, we just need one of these variables in our analysis. Here I keep ‘GarageCars’ since its correlation with ‘SalePrice’ is higher.
* ‘TotalBsmtSF’ and ‘1stFlrSF’ also seems twin brothers. Here I keep ‘TotalBsmtSF’ since it’s correlation with ‘SalePrice’ is higher.
* ‘TotRmsAbvGrd’ and ‘GrLivArea’ twin brothers again.

2.3.2.3 Scatter plots between the most correlated variables (move like Jagger style)





Although I already know some of the main figures, this mega scatter plot gives me a reasonable idea about variables relationships.

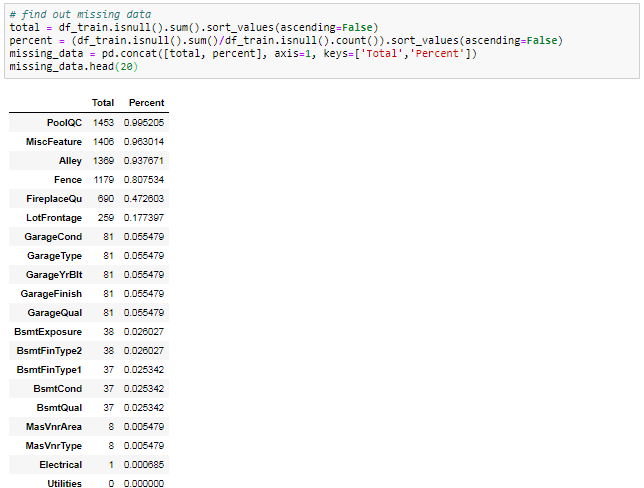
One of the figures we may find interesting is the one between ‘TotalBsmtSF’ and ‘GrLivArea’. In above figure we can see the dots drawing a linear line, which almost acts like a border. It totally makes sense that the majority of the dots stay below that line. Basement areas can be equal to the above ground living area, but it is not expected a basement area bigger than the above ground living area.

1. Missing data
   1. Dealing with missing data

Important questions when thinking about missing data:

* How prevalent is the missing data?
* Is missing data random or does it have a pattern?

The answer to these questions is important for practical reasons because missing data can imply a reduction of the sample size. This can prevent us from proceeding with the analysis. Moreover, from a substantive perspective, we need to ensure that the missing data process is not biased and hiding an inconvenient truth.

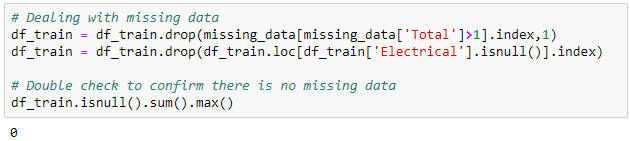


We’ll consider that when more than 15% of the data is missing, we should delete the corresponding variable and pretend it never existed. This means that we will not try any trick to fill the missing data in these cases. According to this, there is a set of variables (like ‘PoolQC’, ‘MiscFeature’, ‘Alley’ and etc.) that we should delete. The point is: will we miss these data? I don’t think so. None of these variables seem to be very important, since most of them are not aspects in which we think about when buying a house (maybe that’s the reason why data is missing?). Moreover, looking closer at the variables, we could say that variables like ‘PoolQC’, ’MiscFeature’ and ‘FireplaceQu’ are strong candidates for outliers, so we’ll be happy to delete them.

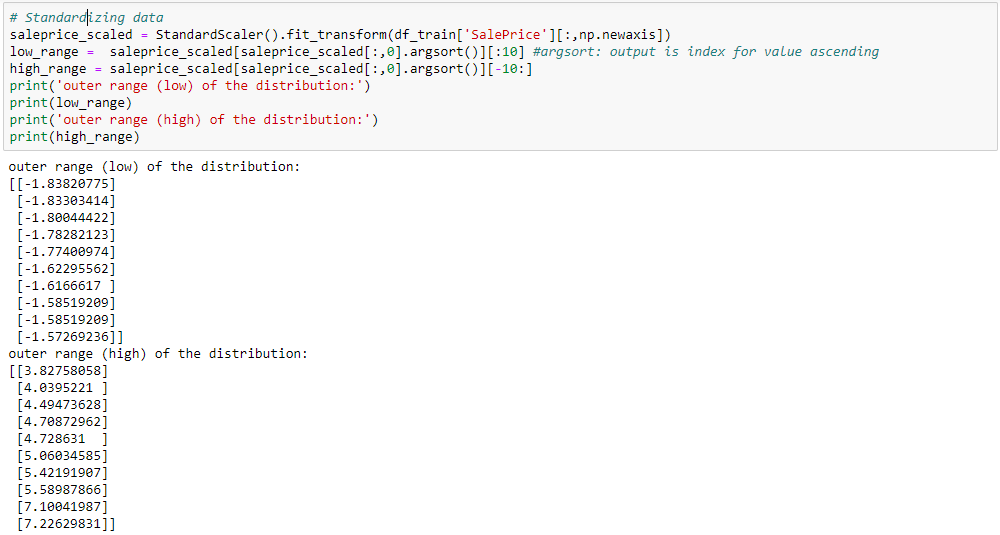
For ‘GarageX’ variables have the same number of missing data. Since the most important information regarding garages is expressed by ‘GarageCars’ and considering that we are just talking about 5% of missing data. I’ll delete the mentioned ‘GarageX’ variables. The same logic applies to ‘BsmtX’ variables.

Finally, we have 1 missing observation in ‘Electrical’. Since it’s just one missing, we’ll delete this observation and keep the variable.

In summary, to handle missing data, we’ll delete all the variables with missing data, except the variable ‘Electrical’. In ‘Electrical’, we’ll just delete the observation with missing data.



* 1. Standardizing data



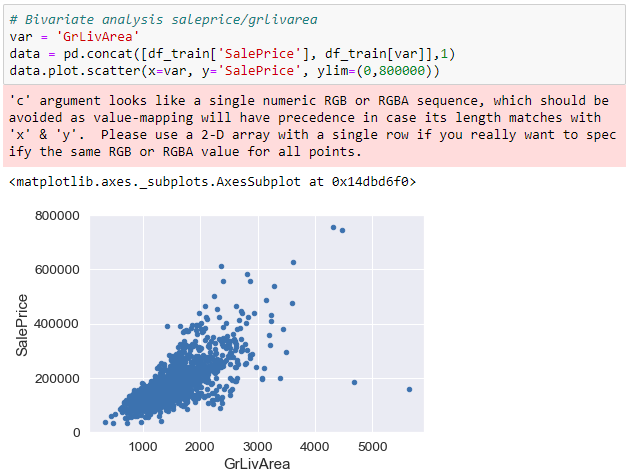
After standardizing, there are 2 finding:

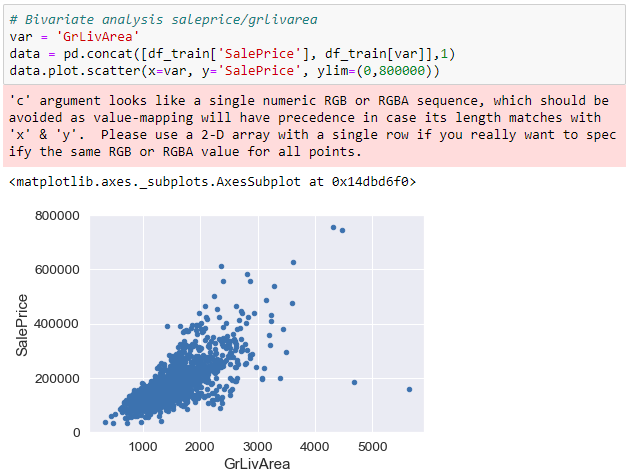
* Low range values are similar and not too far from 0
* High range values are far from 0 and the 7. Something values are really out of range

For now, we’ll not consider any of these values as an outlier, but we should be careful with those two 7. Something values.

* 1. Bivariate analysis

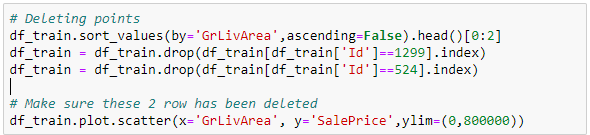
We already know the following scatter plots by heart. However, when we look to things from a new perspective, there’s always something to discover.

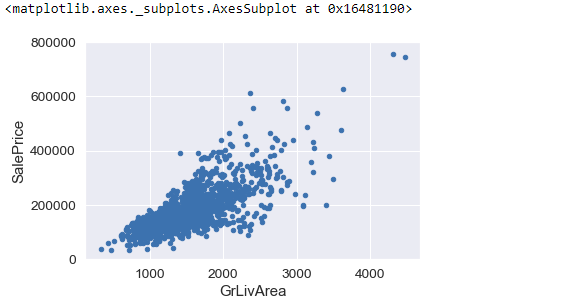


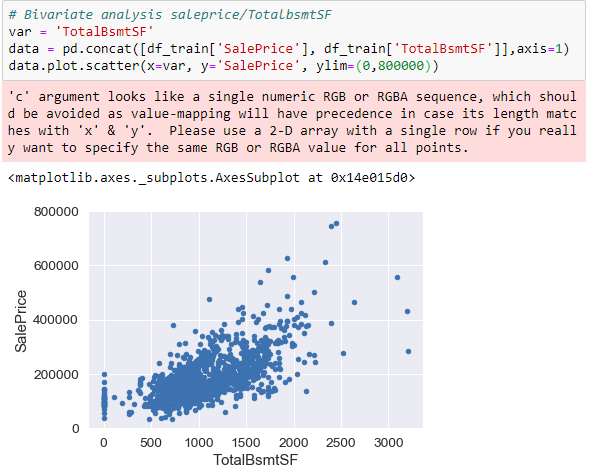


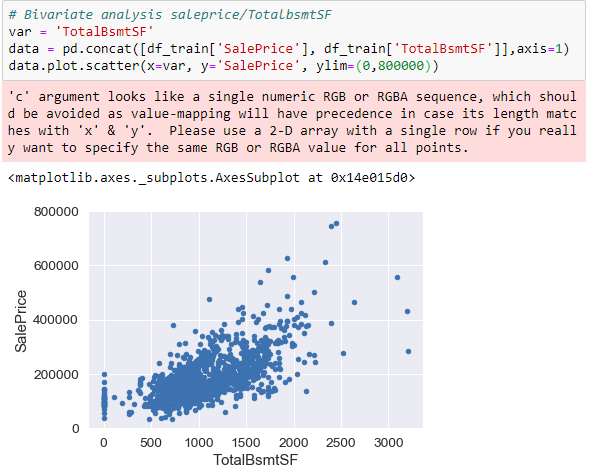
What has been revealed from above scatter plot:

* The two values with bigger ‘GrLivArea’ seem strange and they are not following the crowd. We can speculate why this is happening. Maybe they refer to agricultural area and that could explain the low price. I’m not sure about this but I’m quite confident that these two points are not representative of the typical case. Therefore, we’ll define them as outliers and delete them.
* The two observations in the top of the plot are those 7. Something observations that we said we should be careful about. They look like two special cases. However, they seem to be following the trend. For that reason, we will keep them.









We can feel tempted to eliminate some observations (e.g. TotalBsmtSF>3000), but I suppose it’s not worth it. We can live with that, so we’ll not do anything.

1. Getting hard core

We already did some data cleaning and discovered a lot about ‘SalePrice’. Now it’s time to go deep and understand how ‘SalePrice’ complies with the statistical assumptions that enables us to apply multivariate techniques.

* 1. Model selection

There are several options to choose from when it comes to models. So a good starting point is logistic regression.

* + 1. Logistic Regression
    2. Random forests
    3. Support vector machines
    4. Gradient boosting classifier
    5. K-nearest neighbors
    6. Gaussian Naïve Bayes

1. Evaluation

Now I am going to evaluate model performance and feature importance

* 1. Model performance

We can evaluate the accuracy of the model by using the validation set where we know the actual outcome. This dataset have not been used for training the model, so it’s completely new to the model.

We then compare this accuracy score with the accuracy when using the model on the training data. If the difference between these are significant this is an indication of over-fitting. We try to avoid this because it means the model will not generalize well to new data and is expected to perform poorly.

From 4.1, comparing the score for these models, I though gradient boosting classifier is the best model. It has the highest score for validation data, also the difference between training data and validation data is not big, that means the model does not have over-fitting problems.

* 1. Feature importance – selecting the optimal features in the model

有待补充

1. Deployment

Publishing the resulting prediction from the model to the Kaggle leaderboard.