Assignment Report

Objective

This project is to preprocess and clean the data using different strategies.

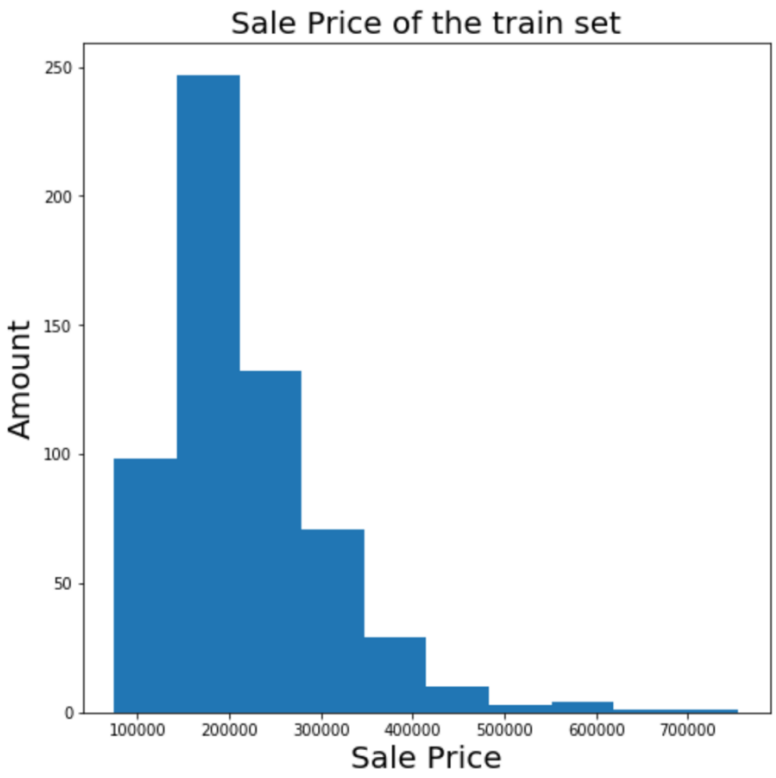
From the data set, there are totally 81 attributes. They can be grouped into three types, nominal, ordinal and numeric. They are grouped as follow:

Nominal: [Id], [MSSubClass], [MSZoning], [Street], [Alley], [LandContour], [LotConfig], [Neighborhood], [Condition1], [Condition2], [BldgType], [HouseStyle], [RoofStyle], [RoofMatl], [Exterior1st], [Exterior2nd], [MasVnrType], [Foundation], [Heating], [CentralAir], [Electrical], [GarageType], [MiscFeature], [SaleType], [SaleCondition]

Ordinal : [OverallQual], [OverallCond], [ExterQual], [ExterCond], [BsmtQual], [BsmtCond], [BsmtExposure], [BsmtFinType1], [BsmtFinType2], [HeatingQC], [KitchenQual], [FireplaceQu], [GarageQual],[GarageCond], [PoolQC], [Fence], [LotShape], [LandSlope], [Functional], [GarageFinish], [PavedDrive], [Utilities]

Numeric: [LotFrontage], [LotArea], [YearBuilt], [YearRemodAdd], [MasVnrArea], [BsmtFinSF1], [BsmtFinSF2], [BsmtUnfSF], [TotalBsmtSF], [1stFlrSF], [2ndFlrSF], [LowQualFinSF], [GrLivArea], [GarageYrBlt], [GarageArea], [WoodDeckSF], [OpenPorchSF], [EnclosedPorch], [3SsnPorch], [ScreenPorch], [PoolArea], [MiscVal], [MoSold], [YrSold], [SalePrice], [BsmtFullBath], [BsmtHalfBath], [FullBath], [HalfBath], [BedroomAbvGr], [KitchenAbvGr], [GarageCars], [TotRmsAbvGrd], [Fireplaces]

The histogram for “SalePrice”:



Data cleaning:

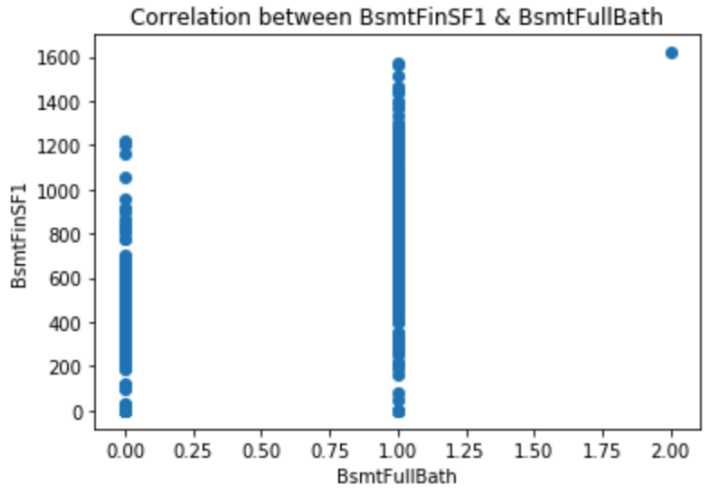
From the data set, there are some outliers in the attributes. 841 records are deleted and 619 records remaining.

There are some attributes that has same value for all records. 7 attributes are deleted. They are [BsmtFinSF2], [LowQualFinSF], [EnclosedPorch], [3SsnPorch], [ScreenPorch], [PoolArea] and [MiscVal].

From the remaining numeric attributes, the correlation coefficient is computed. The top 5 attributes that are most correlated with the attribute, [BsmtFinSF1]: [BsmtFullBath] (0.6492), [TotalBsmtSF] (0.5223), [1stFlrSF] (0.4458), [GarageArea] (0.2967), [BsmtUnfSF] (-0.4952).

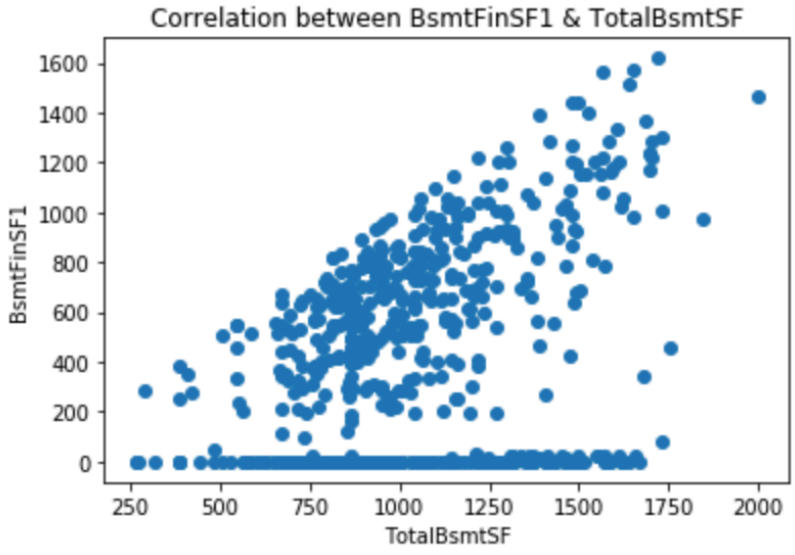
Two attributes are removed. They are [BsmtFullBath] and [TotalBsmtSF].

For [BsmtFullBath]:



From the scatter plot, there are two vertical lines and one outlier. This is graph is overplotting and all the data are derived from the x-axis. There are so few values that [BsmtFullBath] are really categorical scale being represent using number. This is difficult to see the full quantity of values in the dataset.Besides, this has the highest correlation coefficient with [BsmtFinSF1]. This mean that [BsmtFullBath] is highly correlated [BsmtFinSF1], which may lead to redundancy to the data set. So [BsmtFullBath] should be removed.

For [TotalBsmtSF]:



From the scatter plot, this is the second-high correlation coefficient among the five attributes. Also, it is quite linear, which mean that it depends on [BsmtFinSF1], which may lead to redundancy to the data set. So [BsmtFullBath] should be removed.

[GarageCond], [BldgType] and [Alley] are dependent on [GarageQual].

For [GarageQual]& [GarageCond]:

	Ex	Fa	Gd	Po	Ta	row_all
Ex	2	0	0	0	1	3
Fa	0	20	0	4	24	48
Gd	0	0	4	0	10	14
Po	0	0	0	3	0	3
Ta	0	15	5	0	1291	1311
col_all	2	35	9	7	1326	1379

X^2 = (2-3*1/1379)^2/(3*1/1379)+(0-3*35/1379)^2/(3*35/1379)+…

= 2052.5019
Critical value of df 16 = 39.252
2052.5019 > 39.252

So, reject the null hypothesis. [GarageQual] and [GarageCond] are dependent. [GarageCond] should be removed.

For [GarageQual] and [BldgType]:

	1Fam	2fmCon	Duplex	Twnhs	TwnhsE	row_all
Ex	3	0	0	0	0	3
Fa	46	1	1	0	0	48
Gd	11	2	0	0	1	14
Po	2	1	0	0	0	3
Ta	1104	18	39	38	112	1311
col_all	1166	22	40	38	113	1379

X^2 = (3-3*1166/1379)^2/(3*1166/1379) +

(0-3*22/1379)^2/(3*22/1379)+…

= 41.922

Critical value of df 16 = 39.2523
41.922 > 39.2523

So, reject the null hypothesis. [GarageQual] and [BldgType] are dependent.[BldgType] should be removed.

For [GarageQual] & [Alley]:

Alley	Grvl	Pave	All
GarageQual			
Fa	9	1	10
Gd	1	1	2
Po	1	0	1
TA	32	37	69
All	43	39	82

X^2 = (9-10*43/82)^2/(10*43/82) + (1-10*39/82)^2/(10*39/82) +…

= 131.237

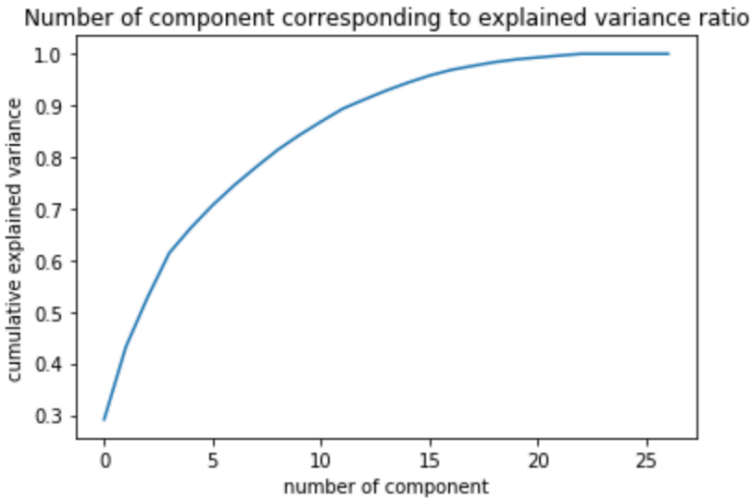
Critical value of df 4 = 18.46682
131.237> 18.46682

So, reject the null hypothesis. [GarageQual] and [Alley] are dependent. [Alley] should be removed.

‘Nan’ in the following attributes are filled with their mean:

LotFrontage: 67.546
LotArea: 9029.256865912763
YearBuilt: 1981.421647819063
YearRemodAdd: 1989.1211631663973
MasVnrArea: 72.82247557003258
BsmtFinSF1: 436.5525040387722
BsmtUnfSF: 603.096930533118
1stFlrSF: 1099.6720516962844
2ndFlrSF: 311.8546042003231
GrLivArea: 1411.5266558966075
FullBath: 1.5783521809369951
HalfBath: 0.3861066235864297
BedroomAbvGr: 2.7722132471728593
TotRmsAbvGrd: 6.24232633279483
Fireplaces: 0.518578352180937
GarageYrBlt: 1985.1099830795263
GarageCars: 1.7867528271405493
GarageArea: 469.5831987075929
WoodDeckSF: 86.58158319870759
OpenPorchSF: 40.508885298869146
MoSold: 6.345718901453958
YrSold: 2007.7883683360258
SalePrice: 175708.5379644588

There are some ‘nan’ value in [MasVnrType] and [Electrical].
For [MasVnrType]: ‘nan’ is filled with the most popular value, ‘BrkFace’.
For [Electrical]: ‘nan’ is filled with ‘SBrkr’.



From the curve, the smallest set of pca feature is 13 when the explained variance is at least 0.9.

The five-number summary of each component is as follow:

```
                0             1             2             3             4   \
count  6.190000e+02  6.190000e+02  6.190000e+02  6.190000e+02  6.190000e+02
mean  -7.855859e-17  7.891731e-18 -1.721832e-17  1.076145e-17 -3.838251e-17
std    2.655975e+00  1.727164e+00  1.451980e+00  1.331233e+00  1.107569e+00
min   -6.758376e+00 -4.044663e+00 -4.037449e+00 -2.708175e+00 -3.256738e+00
25%   -2.186307e+00 -1.355214e+00 -1.034921e+00 -9.586052e-01 -8.167906e-01
50%    2.791552e-01  1.853367e-01 -7.281518e-02 -7.787591e-02 -2.694419e-02
75%    1.818148e+00  1.200873e+00  9.755101e-01  8.141243e-01  7.668494e-01
max    7.791619e+00  5.684919e+00  4.862369e+00  3.720232e+00  3.327879e+00

                5             6             7             8             9   \
count  6.190000e+02  6.190000e+02  6.190000e+02  6.190000e+02  6.190000e+02
mean  -2.152290e-17  1.883254e-18  1.088700e-16 -3.268791e-17 -5.093753e-17
std    1.040126e+00  9.569936e-01  9.317107e-01  9.126736e-01  8.425572e-01
min   -2.603436e+00 -2.257025e+00 -2.525010e+00 -3.080785e+00 -2.666144e+00
25%   -7.107165e-01 -6.571252e-01 -6.045871e-01 -6.381530e-01 -5.525092e-01
50%   -3.633625e-02 -6.874859e-02 -2.893056e-02  9.368013e-03 -1.502531e-02
75%    6.118622e-01  4.918781e-01  6.156252e-01  5.808917e-01  5.788313e-01
max    3.353982e+00  3.324785e+00  3.707755e+00  3.274126e+00  3.056934e+00

               10            11            12
count  6.190000e+02  6.190000e+02  6.190000e+02
mean  -4.143159e-17  1.022338e-17 -4.919440e-17
std    7.985951e-01  7.896067e-01  6.555132e-01
min   -2.734987e+00 -2.275177e+00 -2.422184e+00
25%   -5.159562e-01 -4.824014e-01 -3.533982e-01
50%    2.914667e-02  2.461667e-02  6.071316e-03
75%    5.411788e-01  5.039870e-01  3.714744e-01
max    2.825481e+00  2.708616e+00  2.872221e+00
```

Conclusion

Data cleaning is an important step before analyzing any data. From this report, we first check whether there are some outliers form the numerical attributes. This is because the outliers and extreme cases may affect the result and the accuracy of the analysis. Therefore, we need to drop all the outliers before the analysis. Second, we remove any ‘nan’ value or duplicated records from the dataset. It is because many predictive models can only read the numerical value. Finally, we perform principal component analysis (PCA). This is a method that rotates the dataset in a way such that the rotated features are statistically uncorrelated. This rotation is often followed by selecting only a subset of the new features, according to how important they are for explaining the data, which is an important dimension reduction method to remove any redundant attributes from the dataset.