house prices

1. 认识数据
2. traindata shape：（1640，81）
3. traindata.isnull.sum()

train\_data['PoolQC'].value\_counts(dropna=False)

PoolQC 1453 游泳池质量­——无游泳池

MiscFeature 1406 其他类别未涵盖的杂项功能——无

Alley 1369 小巷通道的类型——无小巷通道

Fence 1179 篱笆质量——无篱笆

FireplaceQu 690 壁炉质量——无壁炉

LotFrontage 259 房产临街距离（线性英尺）——随机森林补全?待考虑

GarageCond 81 车库条件——无车库

GarageType 81 车库位置——无车库

GarageYrBlt 81 车库建成年份——无车库

GarageFinish 81 车库的内部装饰——无车库

GarageQual 81 车库质量——无车库

BsmtExposure 38 地下室墙采光——无地下室

BsmtFinType2 38 地下室1评级——无地下室

BsmtFinType1 37 地下室2评级——无地下室

BsmtCond 37 地下室的一般条件——无地下室

BsmtQual 37 地下室高度——无地下室

MasVnrArea 8 砖石贴面面积——无贴面

MasVnrType 8 砖石贴面类型——无贴面

Electrical 1 电力系统——缺失值，考虑删除

对于“无”的进行缺失值填补，“None”

1. train\_data.describe()

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Id | MSSubClass | LotFrontage | LotArea | OverallQual |
| count | 1460.000000 | 1460.000000 | 1201.000000 | 1460.000000 | 1460.000000 |
| mean | 730.500000 | 56.897260 | 70.049958 | 10516.828082 | 6.099315 |
| std | 421.610009 | 42.300571 | 24.284752 | 9981.264932 | 1.382997 |
| min | 1.000000 | 20.000000 | 21.000000 | 1300.000000 | 1.000000 |
| 25% | 365.750000 | 20.000000 | 59.000000 | 7553.500000 | 5.000000 |
| 50% | 730.500000 | 50.000000 | 69.000000 | 9478.500000 | 6.000000 |
| 75% | 1095.250000 | 70.000000 | 80.000000 | 11601.500000 | 7.000000 |
| max | 1460.000000 | 190.000000 | 313.000000 | 215245.000000 | 10.000000 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | OverallCond | YearBuilt | YearRemodAdd | MasVnrArea | BsmtFinSF1 |
| count | 1460.000000 | 1460.000000 | 1460.000000 | 1452.000000 | 1460.000000 |
| mean | 5.575342 | 1971.267808 | 1984.865753 | 103.685262 | 443.639726 |
| std | 1.112799 | 30.202904 | 20.645407 | 181.066207 | 456.098091 |
| min | 1.000000 | 1872.000000 | 1950.000000 | 0.000000 | 0.000000 |
| 25% | 5.000000 | 1954.000000 | 1967.000000 | 0.000000 | 0.000000 |
| 50% | 5.000000 | 1973.000000 | 1994.000000 | 0.000000 | 383.500000 |
| 75% | 6.000000 | 2000.000000 | 2004.000000 | 166.000000 | 712.250000 |
| max | 9.000000 | 2010.000000 | 2010.000000 | 1600.000000 | 5644.000000 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | BsmtFinSF2 | BsmtUnfSF | TotalBsmtSF | 1stFlrSF | 2ndFlrSF |
| count | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 |
| mean | 46.549315 | 567.240411 | 1057.429452 | 1162.626712 | 346.992466 |
| std | 161.319273 | 441.866955 | 438.705324 | 386.587738 | 436.528436 |
| min | 0.000000 | 0.000000 | 0.000000 | 334.000000 | 0.000000 |
| 25% | 0.000000 | 223.000000 | 795.750000 | 882.000000 | 0.000000 |
| 50% | 0.000000 | 477.500000 | 991.500000 | 1087.000000 | 0.000000 |
| 75% | 0.000000 | 808.000000 | 1298.250000 | 1391.250000 | 728.000000 |
| max | 1474.000000 | 2336.000000 | 6110.000000 | 4692.000000 | 2065.000000 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | LowQualFinSF | GrLivArea | BsmtFullBath | BsmtHalfBath | FullBath |
| count | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 |
| mean | 5.844521 | 1515.463699 | 0.425342 | 0.057534 | 1.565068 |
| std | 48.623081 | 525.480383 | 0.518911 | 0.238753 | 0.550916 |
| min | 0.000000 | 334.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 0.000000 | 1129.500000 | 0.000000 | 0.000000 | 1.000000 |
| 50% | 0.000000 | 1464.000000 | 0.000000 | 0.000000 | 2.000000 |
| 75% | 0.000000 | 1776.750000 | 1.000000 | 0.000000 | 2.000000 |
| max | 572.000000 | 5642.000000 | 3.000000 | 2.000000 | 3.000000 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | HalfBath | BedroomAbvGr | KitchenAbvGr | TotRmsAbvGrd | Fireplaces |
| count | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 |
| mean | 0.382877 | 2.866438 | 1.046575 | 6.517808 | 0.613014 |
| std | 0.502885 | 0.815778 | 0.220338 | 1.625393 | 0.644666 |
| min | 0.000000 | 0.000000 | 0.000000 | 2.000000 | 0.000000 |
| 25% | 0.000000 | 2.000000 | 1.000000 | 5.000000 | 0.000000 |
| 50% | 0.000000 | 3.000000 | 1.000000 | 6.000000 | 1.000000 |
| 75% | 1.000000 | 3.000000 | 1.000000 | 7.000000 | 1.000000 |
| max | 2.000000 | 8.000000 | 3.000000 | 14.000000 | 3.000000 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | GarageYrBlt | GarageCars | GarageArea | WoodDeckSF | OpenPorchSF |
| count | 1379.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 |
| mean | 1978.506164 | 1.767123 | 472.980137 | 94.244521 | 46.660274 |
| std | 24.689725 | 0.747315 | 213.804841 | 125.338794 | 66.256028 |
| min | 1900.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 1961.000000 | 1.000000 | 334.500000 | 0.000000 | 0.000000 |
| 50% | 1980.000000 | 2.000000 | 480.000000 | 0.000000 | 25.000000 |
| 75% | 2002.000000 | 2.000000 | 576.000000 | 168.000000 | 68.000000 |
| max | 2010.000000 | 4.000000 | 1418.000000 | 857.000000 | 547.000000 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | EnclosedPorch | 3SsnPorch | ScreenPorch | PoolArea | MiscVal |
| count | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 |
| mean | 21.954110 | 3.409589 | 15.060959 | 2.758904 | 43.489041 |
| std | 61.119149 | 29.317331 | 55.757415 | 40.177307 | 496.123024 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 50% | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 75% | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| max | 552.000000 | 508.000000 | 480.000000 | 738.000000 | 15500.000000 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | MoSold | YrSold | SalePrice |
| count | 1460.000000 | 1460.000000 | 1460.000000 |
| mean | 6.321918 | 2007.815753 | 180921.195890 |
| std | 2.703626 | 1.328095 | 79442.502883 |
| min | 1.000000 | 2006.000000 | 34900.000000 |
| 25% | 5.000000 | 2007.000000 | 129975.000000 |
| 50% | 6.000000 | 2008.000000 | 163000.000000 |
| 75% | 8.000000 | 2009.000000 | 214000.000000 |
| max | 12.000000 | 2010.000000 | 755000.000000 |

1. testdata: (1459, 80) testdata也存在缺失值，有些字段在traindata没有缺失值但在testdata有缺失值，（应先进行traindata的极端值分析处理，然后再合并traindata和testdata，对缺失值进行填充）
2. 对数值型字段画图分析异常值
3. full\_data.shape (2919, 81)
4. 缺失值

PoolQC 2909 游泳池质量­——无游泳池

MiscFeature 2814 其他类别未涵盖的杂项功能——无

Alley 2721 小巷通道的类型——无小巷通道

Fence 2348 篱笆质量——无篱笆

SalePrice 1459

FireplaceQu 1420

LotFrontage 486

GarageQual 159

GarageCond 159

GarageFinish 159

GarageYrBlt 159

GarageType 157

BsmtExposure 82

BsmtCond 82

BsmtQual 81

BsmtFinType2 80

BsmtFinType1 79

MasVnrType 24

MasVnrArea 23

MSZoning 4

BsmtFullBath 2

BsmtHalfBath 2

Utilities 2

Functional 2

Electrical 1

BsmtUnfSF 1

Exterior1st 1

Exterior2nd 1

TotalBsmtSF 1

GarageCars 1

BsmtFinSF2 1

BsmtFinSF1 1

KitchenQual 1

SaleType 1

GarageArea 1

MSZoning

PoolQC 1453 游泳池质量­——无游泳池

MiscFeature 1406 其他类别未涵盖的杂项功能——无

Alley 1369 小巷通道的类型——无小巷通道

Fence 1179 篱笆质量——无篱笆

FireplaceQu 690 壁炉质量——无壁炉

LotFrontage 259 房产临街距离（线性英尺）——随机森林补全?待考虑

GarageCond 81 车库条件——无车库

GarageType 81 车库位置——无车库

GarageYrBlt 81 车库建成年份——无车库

GarageFinish 81 车库的内部装饰——无车库

GarageQual 81 车库质量——无车库

BsmtExposure 38 地下室墙采光——无地下室

BsmtFinType2 38 地下室1评级——无地下室

BsmtFinType1 37 地下室2评级——无地下室

BsmtCond 37 地下室的一般条件——无地下室

BsmtQual 37 地下室高度——无地下室

MasVnrArea 8 砖石贴面面积——无贴面

MasVnrType 8 砖石贴面类型——无贴面

Electrical 1 电力系统——缺失值，考虑删除

补缺失值，定性数据转换，降维，模型选择，预测

<https://www.kaggle.com/xirudieyi/house-prices>

<https://zhuanlan.zhihu.com/p/34904202>

<https://www.kaggle.com/massquantity/all-you-need-is-pca-lb-0-11421-top-4>