**Kaggle房价预测**

从去年9月份自学python以来，中间经理各种迷茫，但是依然坚持下来了。到现在开始要找工作了，在这之前，要花点时间把自己学习的知识复习回顾一遍，也要做一些实战项目多多练习，以增加熟练度。本项目打算着重复习 python 数据分析（进阶）相关的知识。

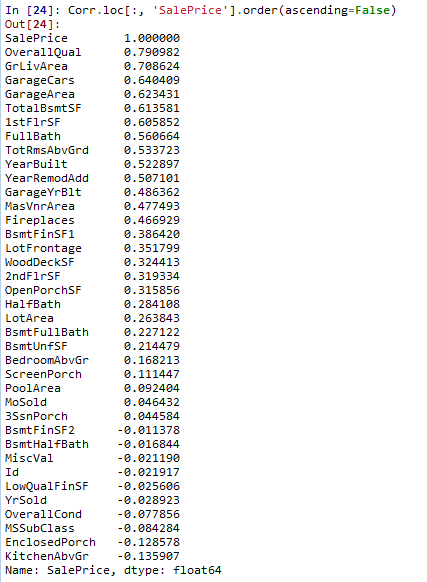
**项目简介：**数据分析是数学统计，计算机科学以及相关业务知识的统一。因此该项目将结合自己房地产工作经验，以及自学的数学统计和分析软件。探索与房地产价值相关的因素。

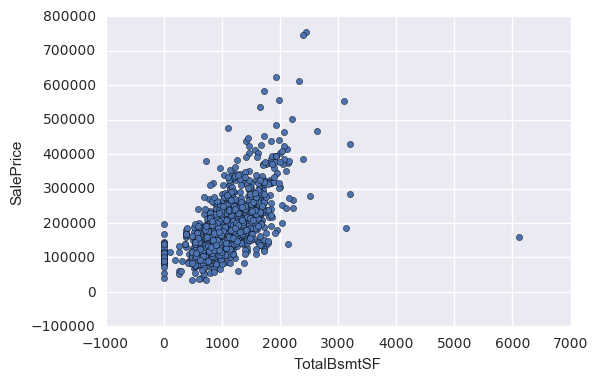
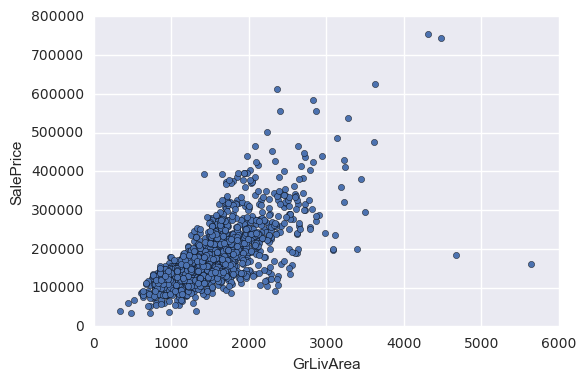
**数据来源：**来自于kaggle竞赛：House Prices: Advanced Regression Techniques的数据集

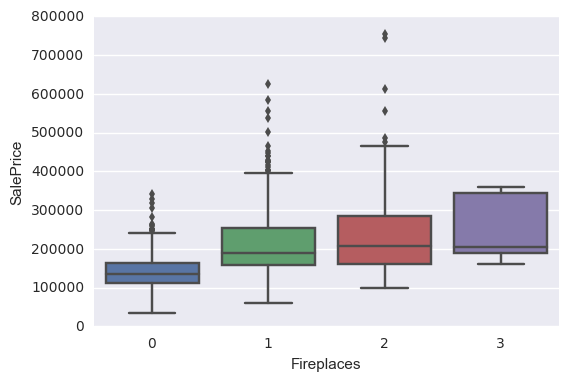
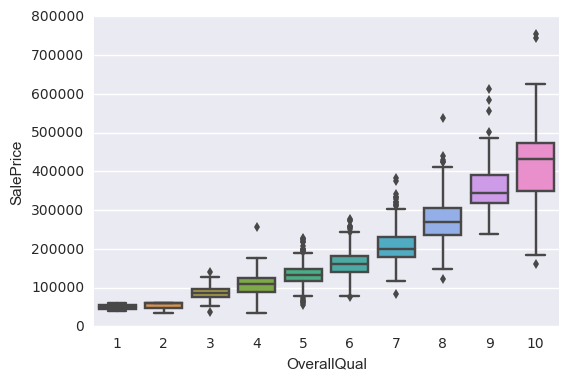
**目的：**探寻与房地产价值相关的因素，然后建立模型，预测房地产价值。

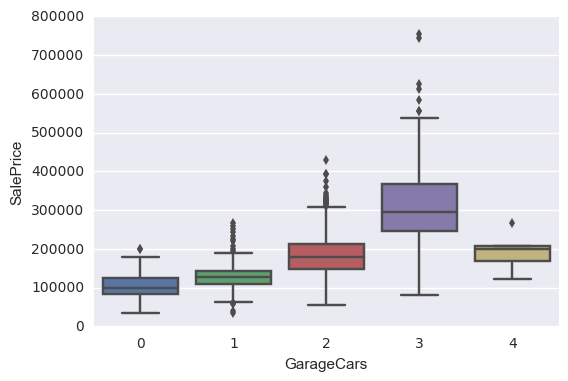
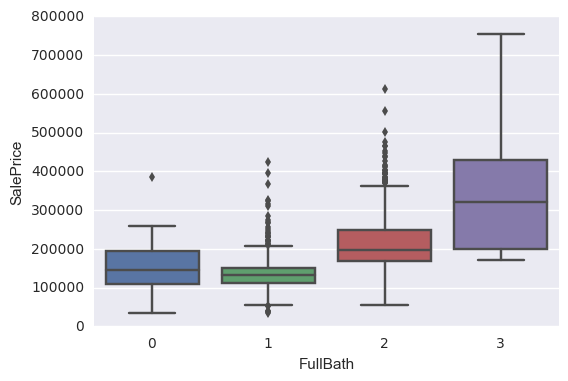
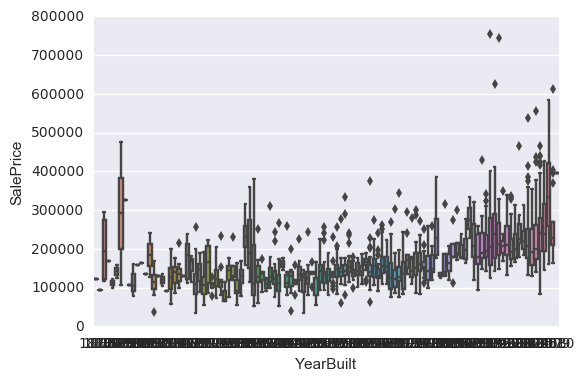
**数据初步探索：**

通过其他数值属性与SalePrice的相关关系，快速寻找与SalePrice相关因素较大的数值属性。

  
在以上属性中，有多个属性从不同维度来描述同一特征。通过比较变量相关关系以及变量与目标相关关系，应使用如下属性：OverallQua（房屋用料及完好程度总体评价），GrLivArea（房屋地上居住面积），GarageCars（车库容纳车数量），TotalBsmtSF（地下室总面积），FullBath（卫生间数量）， YearBuild（建筑年代），Fireplaces为壁炉数量。



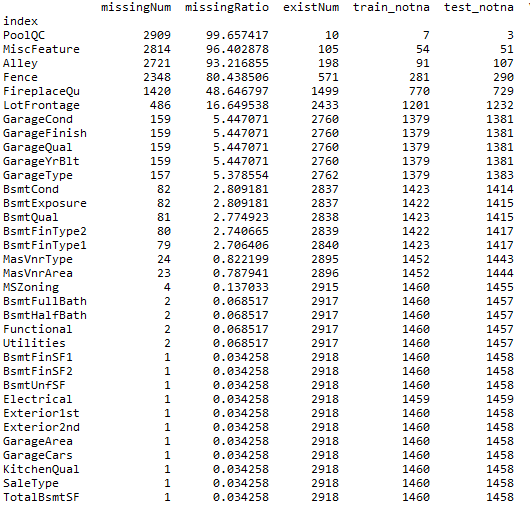


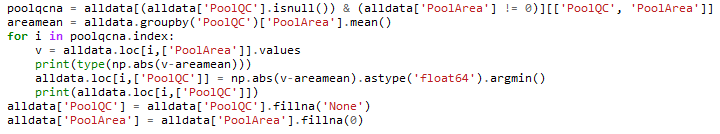
上图为七个属性与SalePrice组成的散点图或箱型图。可以看出随着GrLivArea以及TotalBsmtSF的增大，SalePrice总体趋势是上升的。OverallQua评价越高，则满意度越高，房价也就越高。Fireplaces，FullBath，GarageCars越大，SalePrice趋势也是上升的。房屋越新，使用年限越短，损坏越少，房价越高，当然也有一些特殊情况，如某些房子可能存在特殊的历史意义等。

**处理缺失值：**

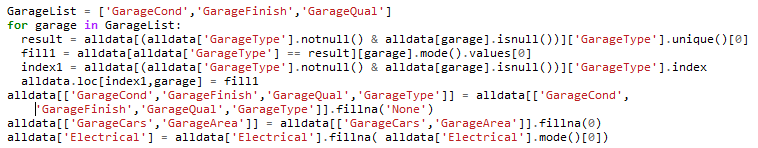
在该数据集中，有部分属性从不同角度描述了不同的物体，因此处理缺失值时，可以从两个属性之间的联系来处理。有部分NA则表示物体不存在，这样的缺失值可直接填充None，或0,（注意：NA既表示缺失值，又表示不存在的情况）



寻找PoolArea不为0但PoolQC为NA的索引，根据面积填充PoolQC，PoolQC和PoolArea其他的Na值用None和0填充



寻找GarageCond不为空而GarageCond,GarageFinish,GarageQual为空的索引，用分组众值填充，其他的均用None或0值填充



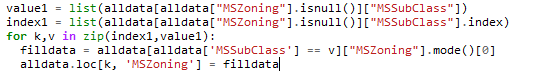
对含有Bsmt的属性进行空缺值进行处理



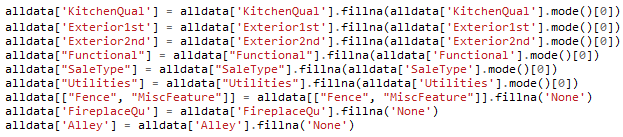
由MasVnrArea和MasVnrType关系，对MasVnrArea不为空和MasVnrType为空值填充一定范围内的众值，其他的填充None或0。



由于MSZoning和MSSubClass存在一定关系，用分组众值填充MSZoning空值

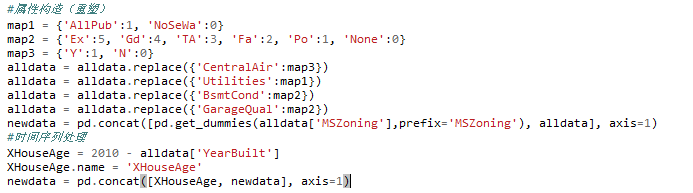


其他的空缺值均由众值填充



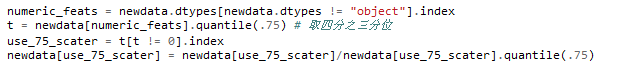
**数据变换**

对CentralAir,Utilities,BsmtCond,GarageQual序列型属性转化为数值序列型数据，将YearBuilt转化为连续型数据，将分类变量MSZoning转换为“哑变量矩阵”，将转变后的数据重新与SalePrice进行相关性检查后，去掉Utilities



**特征缩放：**

对数值型特征进行特征缩放，取3/4位进行缩放。

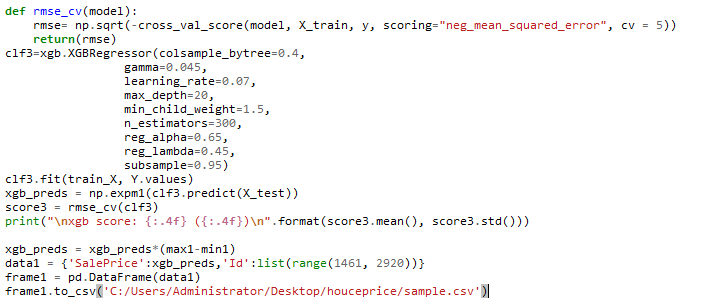


对SalePrice进行处理，使用 x/(xmax-xmin)



**建立模型**

使用xgboost模型中得XGBRegressor进行拟合，得到结果后按要求保存到sample.csv中



**总结：**

本项目的重点（对我来说）在于数据的清洗和数据转换，以及特征值的选取。在特征选取方面，我根据个人经验以及数据分析技术进行了选择过滤，同时对选择的特征进行了初步的分析.特征涵盖了房屋面积，地下室，车库，卫生间等状况，也包括外部环境。严格说来，这样选择并不严谨，或许遗漏了某些特征值。

另外在机器学法的选择上，我也只能进行初步运用，而且使用之后，也无法评论其好坏。这些知识只能在以后慢慢积累。