1、查看不同的变量以及描述性数据

2、绘制SalePrice直方图，发现房价偏离正太分布（正偏），对房价进行对数转换。

3、房价的相关变量分析: 选出自己认为跟房价相关性较高的几个变量

3.1 数据型变量：地块面积，地下室以上生活区面积，画出散点图，发现两个变量和房价都关系密切，基本呈线性关系，并且都是正相关.LotArea斜率更高一些。

3.2 类别型变量：TotRmsAbvGrd, OverallQual, YearBuilt和房价有关,TotRmsAbvGrd, OverallQual的相关性更强,TotRmsAbvGrd值在11附近房价最高,OverallQual越高房价越高. 其他的类别没有趋势性,但是存在不同类别对应房价不同的情况

4、客观分析

4.1 相关系数矩阵（计算所有特征值每两个之间的相关系数，并作图。）

图中不同变量之间，颜色越强的相关性越强，可以看出相关系数最大的有， TotalBsmtSF和1stFlrSF； Garage变量群； YearBuilt和GarageYearBuilt;

和SalePrice相关性较大的有（相关性依次递减）： OverallQual; GrLivArea; TotalBsmtSF、lstFlrSF、GarageCars、GarageArea

4.2 SalePrice相关系数矩阵

从图中可以看出：

1、OverallQual, GrLivArea 以及 TotalBsmtSF 与 SalePrice有很强的相关性。  
2、 GarageCars 和 GarageArea 也是相关性比较强的变量.车库中存储的车的数量是由车库的面积决定的，不需要专门区分GarageCars 和 GarageArea ，所以我们只需要其中的一个变量。这里我们选择了GarageCars因为它与SalePrice的相关性更高一些。  
3、TotalBsmtSF 和 1stFloor 与上述情况相同，选择 TotalBsmtSF。  
4、FullBath几乎不需要考虑。  
5、TotRmsAbvGrd和 GrLivArea与2情况相同，选择，GrLivArea。  
6、YearBuilt 和 SalePrice相关性似乎不强。

4.3 SalePrice和相关变量之间的散点图，丰富的散点图给了我们一个关于变量关系的合理想法。

1、TotalBsmtSF 和 GrLiveArea之间的散点图是很有意思的。  
我们可以看出这幅图中，一些点组成了线，就像边界一样。大部分点都分布在那条线下面，这也是可以解释的。地下室面积和地上居住面积可以相等，但是一般情况下不会希望有一个比地上居住面积还大的地下室。  
2、SalePrice 和YearBuilt 之间的散点图也值得我们思考。在“点云”的底部，我们可以观察到一个几乎呈指数函数的分布。我们也可以看到“点云”的上端也基本呈同样的分布趋势。并且可以注意到，近几年的点有超过这个上端的趋势。

5、缺失数据处理

5.1、当超过15%的数据都缺失的时候，我们应该删掉相关变量且假设该变量并不存在。根据这一条，一系列变量都应该删掉，例如PoolQC, MiscFeature, Alley等等，这些变量都不是很重要，因为他们基本都不是我们买房子时会考虑的因素。  
5.2、Garage 变量群的缺失数据量都相同，由于关于车库的最重要的信息都可以由GarageCars 表达，并且这些数据只占缺失数据的5%，我们也会删除上述的Garage X 变量群。  
5.3、同样的逻辑也适用于 Bsmt X 变量群。（用BsmtFinSF1）  
5.4、对于 MasVnrArea 和 MasVnrType，我们可以认为这些因素并不重要。除此之外，他们和YearBuilt以及 OverallQual都有很强的关联性，而这两个变量我们已经考虑过了。所以删除 MasVnrArea和 MasVnrType并不会丢失信息。  
5.5、最后，由于Electrical中只有一个损失的观察值，所以我们删除这个观察值，但是保留这一变量。

6、异常值处理

6.1 单因素分析 ：对数据进行正态化，意味着把数据值转换成均值为0，方差为1的数据。

6.2 双变量分析

GrLivArea和SalePrice双变量分析

TotalBsmtSF和SalePrice双变量分析

7、核心部分

7.1 正态性

应主要关注以下两点：  
直方图 – 峰度和偏度。

正态概率图 – 数据分布应紧密跟随代表正态分布的对角线。

可以看出，房价分布不是正态的，显示了峰值，正偏度，但是并不跟随对角线。

数据的正太概率图在最大值和最小值有点偏离直线，所以这个属性的左右两边尾部的数据要多于高斯分布的尾部数据，可以用对数变换来解决这个问题。

对其他几个特征进行正态性分析。

8、将类别变量转换为虚拟变量

9、建模

9.1 使用交叉验证去找到模型最佳参数.  
简单来说，交叉验证的好处就是考虑了多种可能之后，  
尽全力找到你的数据最好的划分方式，用这种方式算出最好的参数。

9.2 因为对房价的预测需要预测出来确切的数值，所以首先选择了岭回归方法，训练模型，找到合适的参数（交叉验证），用训练数据训练模型，再用测试数据进行预测，得到最终结果。

之后尝试了RandomForestRegressor方法、xgboost方法，最终的效果最好的是xgboost.

kaggle：

首先查看不同的变量和描述性数据，先对这些变量有个大概的理解。

对变量进行相关性分析，找出来相关性比较强的一些变量。

通过房价的相关系数矩阵，分析出跟房价相关的一些变量，并且对他们进行判断和解释，

会找出不重要的变量，以及类似的变量。

画出房价和选出来的相关性比较强的变量，之间的散点图，对他们之间的关系作出解释。

通过上面的分析，得出相关缺失值和异常值的处理方案。

对不同变量的正态性进行分析，比如房价的直方图，有正偏的趋势，正太概率图的首尾两端偏

离直线，对它进行对数转换。

1. 认识数据，理解数据，含义；主观的认识解释客观的现象
2. 数据预处理：缺失值、异常、标准化的处理
3. 模型选择