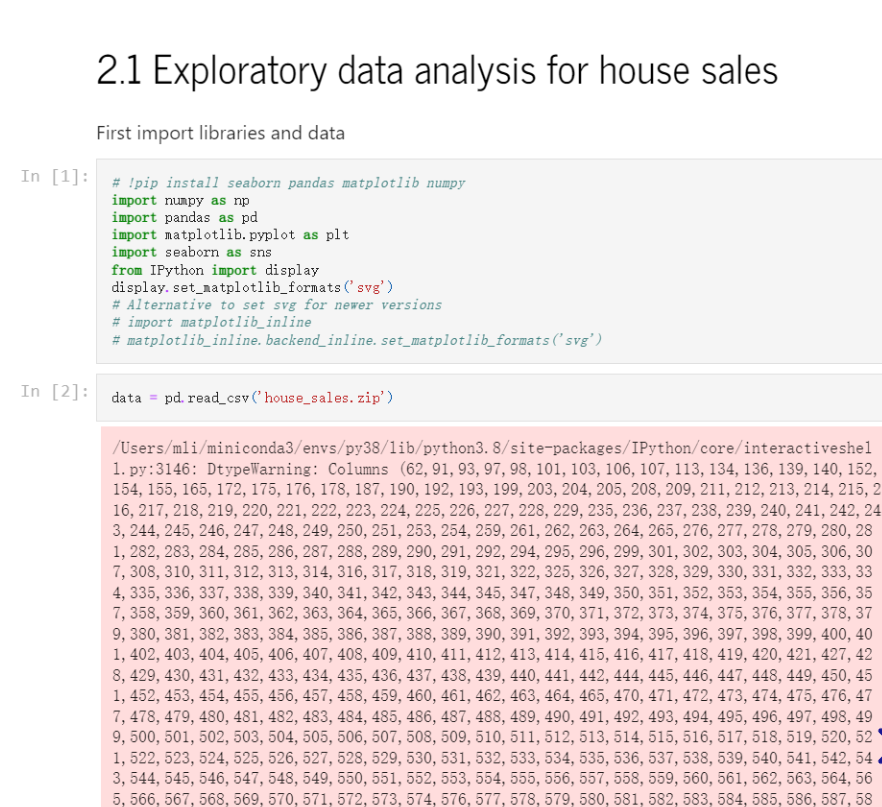
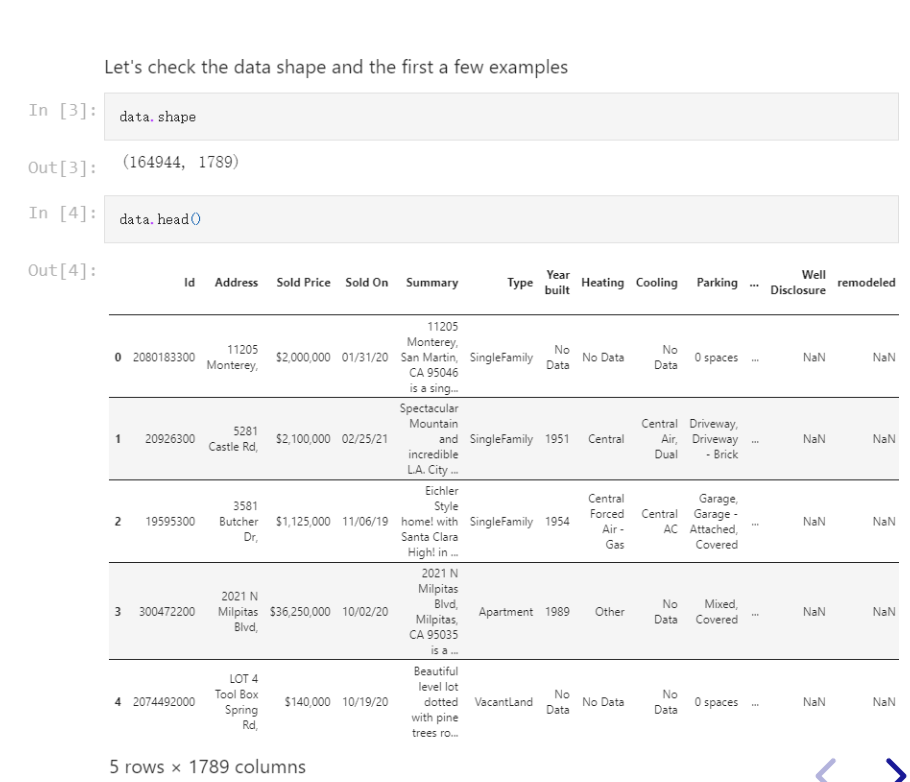
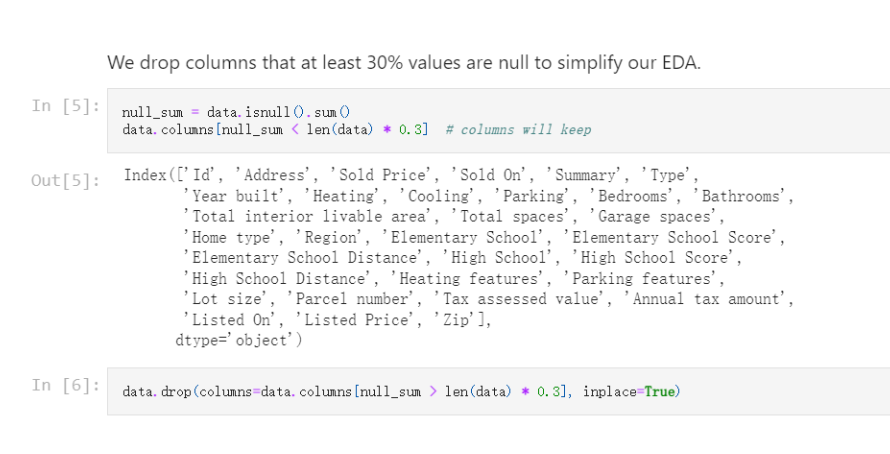
<https://c.d2l.ai/stanford-cs329p/_static/notebooks/cs329p_notebook_eda.slides.html#/>



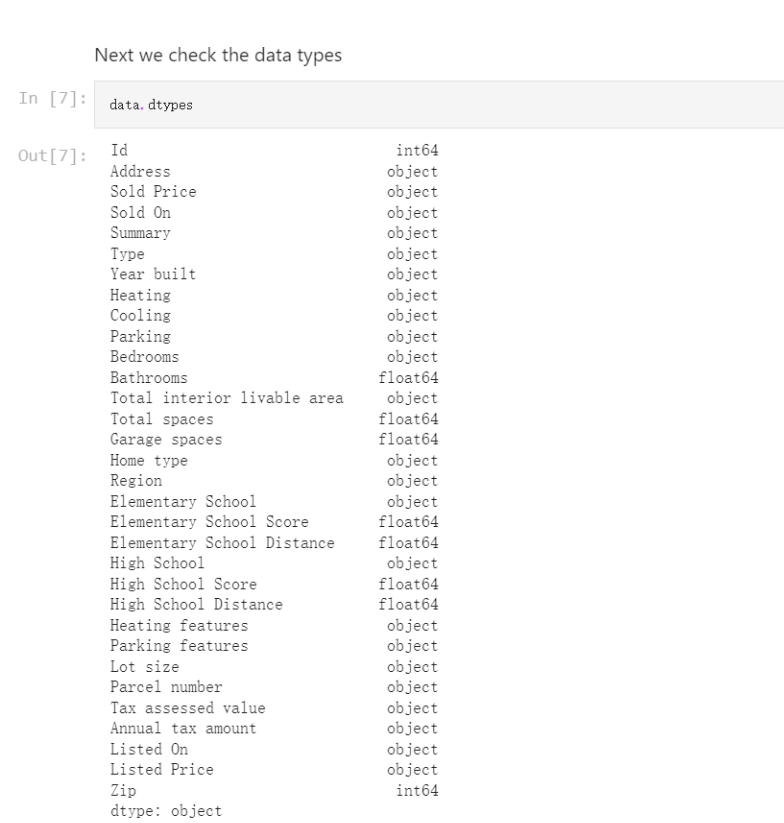
1. 上集我们说的那个卖方的例子，就是从网上爬下来，然后打包下来成house\_sales.zip。
2. Matplotlib就是从matlab借鉴过来的，seaborn就是在matplotlib基础上加了很多画法。
3. 读数据：pd.read\_csv(.zip),csv文件是可以压缩成zip文件，pd可以直接读出zip文件！！图片就没法压缩了，文本是可以压缩的。
4. 这有很多warning，就是python解析不了这些数据，有很多脏数据，这没关系，因为网上爬下来的，奇奇怪怪的数据很多



1. .head()看一下数据，能看见很多没数据，NaN脏数据等等，数据科学家就是为了处理这些数据，这一小节就是大概看一下数据集是什么样的，下面几节会介绍几个常用的技术来处理这些脏数据。



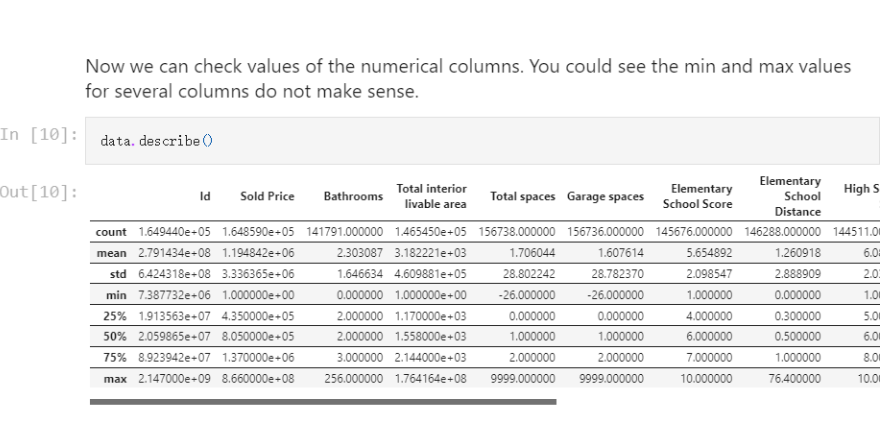
1. Null\_sum就是统计空数据的个数
2. 下面的就是要把缺少30%的数据以上，就把这些列drop掉，注意是inplace，也就是说直接对数据集进行修改，这样省空间，只能跑一次。



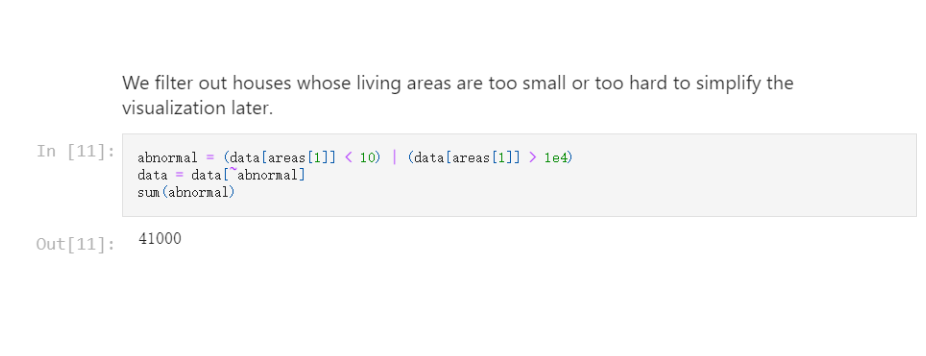
1. 接来下我们看一下我们的type，就比如sold price表示卖出去的价格，应该是一个float类型，year built应该是一个integer等等，原因可能是因为爬下来的数据python无法识别，有一些脏数据，所以这里要做数据转换，这也是数据清洗的一部分。



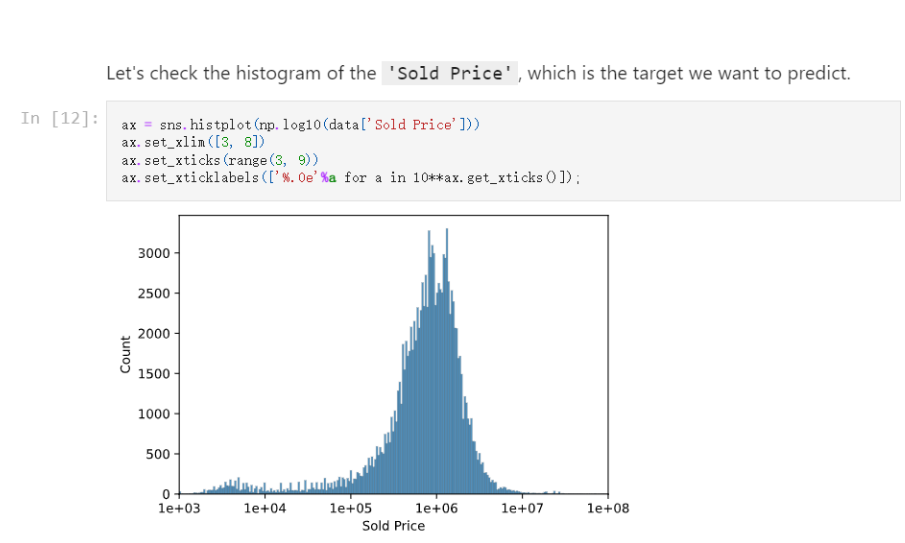
1. 第一个块就是先找出对应的列，然后用正则表达式进行替换，比如美元符号、逗号、横杠都删掉，然后空字符串转成np.nan格式，然后修改格式为float格式。
2. 就比如面积的单位有sqft和acres，就要统一单位，1acres等于43560sqft，就成起来就行，处理一下。



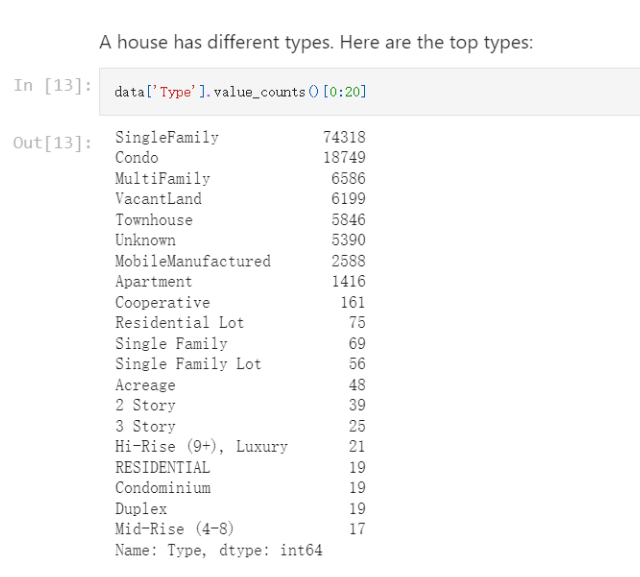
1. Describe一下，每一列的数值，就大概知道每一列的数值是什么，就能大概知道哪里有问题。就比如total space的min是-26，最大是9999，这应该就是不合理的，就比如bathroom最大有256个？？就有些列可能看起来正常，有些列就看起来很不正常，所以需要处理一下



1. 举个例子处理一下，过滤一下居住面积太小或者太大的房子，就是小于10平米，大于1000平米，然后过滤了4w个房子，一共6w个，问题不大。

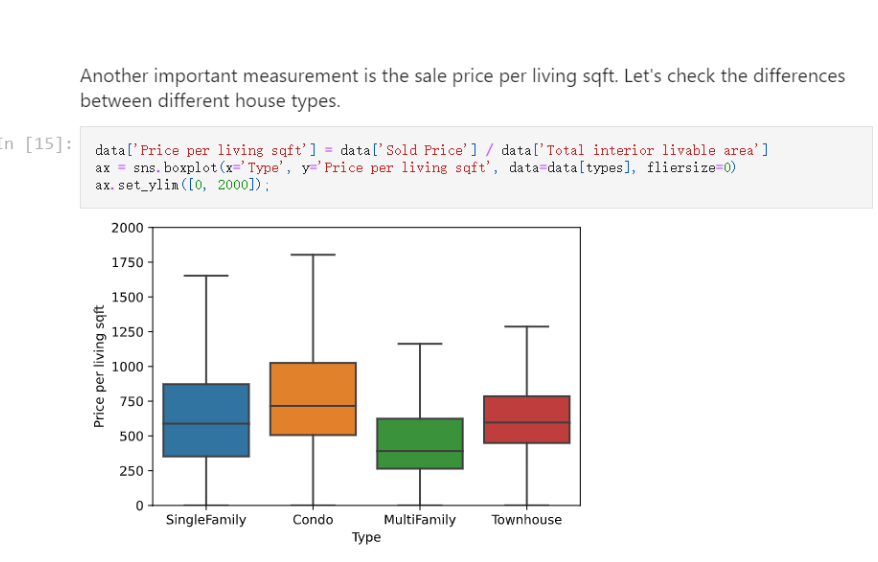


1. 注意用了log10，因为一些价格差距会很大，所以log10以后，均价会比较平稳，然后我们就会看到我们的预测值大概分布是什么样的，看一下均值在100w左右，便宜的在10w，贵的在1000w。



1. 然后我们看一下房子的类别有多少个，singlefamily是7w个，single damily中间加一个空格等等，这里就应该是弄错了，需要在数据清洗的章节中处理好。

  
1. 这个就是看一看房子的类型和价钱的分布密度关系，就是用了核密度估计（kde）



1. 一平米多少钱？单位面积的价格，box plot箱型图



1. 邮政编码（不同位置）的单位均价



1. 协方差，看看向量的相关性，不同的列相关性

