在特征工程这一部分，我们一共对三个变量进行了处理。

第一个处理的变量bathroom\_text。首先我们在不同的room\_type下发现有许多非常繁杂的标签，所以我们将这些标签进行了归类，分别是没有完整的卫生间以及不同的卫生间数目。并且在room\_type为private时，这里还分为私人卫生间和共享卫生间。

第二个进行了归类的特征是property\_type，就是建筑类型。同样将所有的标签列出也十分繁杂，但通过根据在爱彼迎网页上的调查发现，所有的标签共可以分为14种建筑类型，这里列举在了右边，我们将原数据的标签归类到这14种新的标签中。

第三个处理是关于是否可预定的几个变量，我们认为这些变量的大部分信息完全可以用一个新的变量来替代，这个新的变量有五个标签，分别代表这个房源在30天内是可预定的，在30天内不可预定但是60天内可预定，然后是90天，365天以及最后的一个标签表示不可预定。

我们在预处理的第三个部分是情感分析:

我们处理的是两个复杂的文本变量description(1099)和neighborhood\_overview(4368)，我们发现在这两个变量中，相较于在房源描述以及周边概况是否具体提及一些事物，最能体现不同房源之间区别的其实是host对于这间屋子介绍的积极程度，我们希望可以通过情感分析将这两个复杂文本量化转变为0～1之间的数值，来代表host对于这个房源的积极性。

房源描述和周边概括的文本可以分为中文和非中文，分别采用SnowNLP和TextBob中的方法进行情感分析。值得一提的是，SnowNLP返回的情感极性取值变化范围是[0,1]，表示的的是‘这句话代表积极情感的概率’，比较符合我们的期望。但TextBob返回的情感极性的变化范围是[-1,1]，-1代表完全负面，1代表完全正面。所以我们将TextBob返回的情感极性通过一个函数映射投影到[0,1]之间。另外，我们将这两个变量中的缺失值设置为0，这意味着如果host对于房源没有描述或者没有介绍周围情况，我们会选择给他最低的情感评分。

我们预处理的第四个部分是充分降维:

在数据清洗中我们已经将设施中出现频次较多转变为71列哑变量，但是这么多变量对于数据分析来说太过冗长。所以我们希望对这些变量进行降维，同时尽量避免损失关于price的信息。同样根据爱彼迎网页上的调查发现：这些设施大概可以分为五个方面：分别是入住，厨房，洗浴，安全以及基础设施。

在这里我们采用了充分降维的方法：可以通过类比因子分析来理解这个降维方法，但不同的是，为了加入有关price的信息，所有变量和得出的因子都是在因变量price的条件下，最终得出所有观测在这五种类型设施下的得分。并且我们对得分进行了归一化处理，得到的效果如右下角所示。

在预处理后我们的数据集仍有27784个观测，接下来由我的队友来进行数据分析。