**研究题目：**利用KNN对爱彼迎中的房间价格进行预测

**问题描述：**当考虑一个酒店房间的价格时，有多方面因素，如：房间类型、床的数量、可住人数等等。在消费者考虑酒店房价是否合理时，应该先根据已有数据，也就是上述因素，来预估房间的价格，再与实际进行比较。在这种情况中，一个合理的计算房价的算法就十分有必要了。因此，在这篇报告中，我们提出了使用KNN算法预测房价的方式。

**探究过程：**

1. **确定价格的判定和影响因素**

在给定的数据集中，共有三种价格。我们认为，清洁费用和押金与房屋好坏并无直接关系。因此，我们决定直接选用房价。

在给定的数据集中，共有16种因素，其中包括海拔和纬度等无效因素。因此，我们决定先对因素做筛选处理。首先，我们先筛除了数据大量重复，没有区分度的因素，如：海拔、维度、城市和床的数量等。同时，我们认为，当价格从小到大排序，其影响因素应具有大致从小到大或从大到小的顺序，也就是影响因素应与价格大致成正比或成反比。因此，我们筛除了一些没有规律的数据，如：回应率、接受率、最短/长住天数等。

我们最终选用的因素如下：

1. 容纳人数
2. 房型种类
3. 卧室数量
4. 卫生间数量
5. **对于影响因素的处理**

在因素“房型种类”中，所给出的数据并不是以数字量化的形式，而是以“shared”、“private”和“entire”三种文字方式。因此，我们需要对其量化。

通过查阅资料与价格分析，我们决定将“shared”规定为1，“private”规定为3，“entire”规定为10。

1. **对于数据的处理**

在数据集中，有一些无效数据。如，卧室数量所对应数据为“real bed”。我们在Excel中做筛选并删除处理。

同时，数据集中有一些空数据，我们处理的方式是将其规定为0。

1. **距离计算**

采用欧氏距离计算方式，计算公式为

1. **算法思路**

对于任意点，通过欧式距离计算所得结果，找出与其最接近的k个点，并找出其对应的价格。该点价格则被预测为重复最多的对应价格。但是在此数据集中，由于价格重复的概率比较低，则在没有重复价格时，将该点预测值规定为距离最近的点的价格。

准确率的确定：我们将预测价格与实际价格的偏差作为偏差值，除以实际价格得出误差率。再用一减去误差率得出单个点的准确率。最后，我们将所有测试点的准确率取平均值，得出总体的准确率。

1. **K的取值探讨**

我们利用matplotlib对于不同的k取值进行了画图分析。图一中为k的取值范围是3-200，每隔3个取一个；图二中的取值范围为3-2000，同样步长也是3。这样可以避免k变成偶数以至于取不出最优解的情况。我们可以看出大致的k的取值对于准确率的影响。

当k在200-1250之间时，准确率最大，约为68%-70%。K在3-200之间时，k与准确率约为正比关系。当k在1250-1375时，准确率随着k的增大而减小。当k在1375-2000之间时，准确率在58%左右浮动。

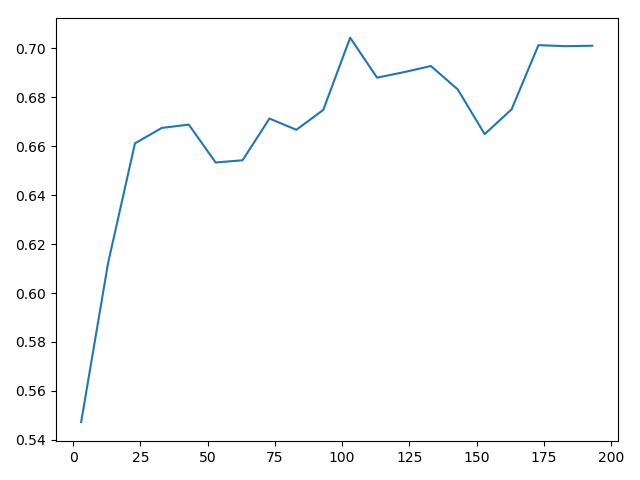


Figure 1



Figure 2

对于k取值的结论：我们认为，综合准确率与所消耗的时间与内存，k取值在200左右比较合适，同时k应该为奇数。