# 业务理解

Airbnb作为一款短租潮流的引领者，成功的吸引了全世界范围内很多产品的模仿，也引发了专业短租公司的兴起。并且，成功的将共享经济的模式从原有的基础行业拓展到了房地产市场。Airbnb通过不断的努力运营，成功的塑造了清新优雅的产品形象（例如，我在研究中发现，在发展初期，部分商户发现照片精美的客房更能吸引用户，故请了摄影师进行拍摄，逐步的形成了产品的品牌形象）这是一个值得分析的商业运营案例

## 结合数据选择目标受众

Airbnb的主要参与者有三类：旅行者，房东和其他体验提供者

旅行者：基本和普通旅游产品用户一致，是进行短期旅行、需要在目的地短期租房的房客。特殊的是，airbnb的旅行者对住宿的需求和认识高于普通旅行者，已经厌烦了跟团游的任务感、强目的性和传统旅馆或酒店的标准化服务，这部分用户开始倾向于自由行的闲适，从原有的探索式旅行（对未见过事物的好奇心）开始转变为体验式旅行（对未感受过的生活方式的好奇心），在从商业分析的角度来看这是一种消费升级在旅游业中的体现。

房东：拥有空闲的房子或房间，希望通过将闲置空间出租赚取利益的群体。而由于Airbnb旅行者用户的较高要求和个人房东的精力有限，现阶段全球范围内产生了一批专注做民宿短租的公司来标准化设计民宿。

体验提供者：

指代提供体验服务的当地居民，他们拥有一些当地的文化特殊技能或熟悉当地特色内容，能够为用户带来当地文化的有趣体验，并以此来获取盈利、结交好友。

显然，选择目标受众需要站在数据的角度，课程提供的数据是应该是从airbnb网站上利用合适的接口爬取的数据，共有99个变量，是针对airbnb的截面数据，因此问题的提出必须站在这些airbnb的拥有者的立场上，即上文中提到的房东。在房东的立场上提出业务问题，首先被我们注意到的应该是如何提高airbnb短租房屋给他们带来的收入。

## 确定具体的业务问题

对这个问题的研究在这里分为两个角度：价格和需求。也就是说，要使一个airbnb短租房屋为房东获得最大收益，这个问题的的答案应该是airbnb价格和旅行者对airbnb的需求综合决定的。虽然在这个过程中可能还有其他的影响因素，但显然作为主要因素的是这两者。

## 如何进一步研究价格问题

研究价格的影响是比较容易的，这是由于数据中清晰的提供了每个airbnb的价格，包括可能产生的各种服务费用。研究其他变量对价格的影响或探求它们之间的关系有许多可以选择的理论与方法。例如，两种可行并即将被本文采用的方法就是：利用回归方法和决策树方法研究价格变量与其他变量间的关联。

这两种方法在本文中被用于于建模和分析变量之间关系，更具提的问题是：价格与其他变量如何结合并且与一起产生特定结果。 由于大部分是数值型变量或可以被转换为数值型变量，因此线性回归方法是比较合适的，本文将使用线性模型对多个独立输入变量（特征变量）和输出因变量之间的关系创建模型。 模型保持线性，输出是输入变量的线性组合。

对于决策树方法，我们选择将价格变量按照区间分为不用的定序变量。由于给定的数据集数据量大，部分数据信息密度比较低，我们在算法上选择CART算法，这是由于该算法可以选择合适的时刻结束。换句话说，理想地说，每一个叶节点里都只有一个类别时分类应该停止，但是很多数据并不容易完全划分，或者完全划分需要很多次分裂，必然造成很长的运行时间，所以CART可以对每个叶节点里的数据分析其均值方差，当方差小于一定值可以终止分裂，以换取计算成本的降低。

## 如何进一步研究需求问题

研究需求的影响则比较困难，主要原因是提供的数据中并不含有很多与需求相关的变量，还有一个重要原因是需求富于弹性的商品。其收益与价格是反方向变动的，即总收益随价格的提高而减少，随价格的减少而降低；需求缺乏知弹性的商品，其销售收益与价格是同方向变动的，即总收益随价格的提高而增加，随价格的降低而减少。

因此，需求本身及它的相关因素是一个很难利用固定数据研究方法得出结果的问题，庆幸的是，针对这种问题，我们有一些方法专用于解决这一类型的问题，例如聚类方法，作为无监督学习的一个重要方法，聚类的思想就是把属性相似的样本归到一类。对于每一个数据点，我们可以把它归到一个特定的类，同时每个类之间的所有数据点在某种程度上有着共性，在算法逻辑中，可能是空间位置接近等特性，在实际业务逻辑中，就是某些变量之间存在趋同性，可以借助算法加以整理，从而突出主要变量的影响。

# 数据概览和预处理

在数据转换阶段，我们对数据进行采样处理、类型转换、归一化。采样是从特定的概率分布中抽取样本点的过程。采样在机器学习中有非常重要的应用：将复杂分布简化为离散的样本点；用重采样可以对样本集进行调整以更好地进行调整并适应后期的模型学习；用于随机模拟以进行复杂模型的近似求解或推理。采样的一个重要作用是处理不均衡数据集。

但在进行这些任务之前，手动选择需要的数据列是一项重要的任务。特殊列的处理方法与缺失值的处理方法也会显著影响分析结果的质量

## 2.1，决定使用哪些列

根据我们选定的分析方法，结合airbnb数据集的具体情况，我们将不具有分析意义的变量，如“ID\_airbnb”，“title\_airbnb”，在进入编程语言之前删去。

## 2.2条。处理特殊列

数据类型可以简单划分为数值型和非数值型。数值型有连续型和离散型。非数值型有类别型和非类别型，其中类别型特征中如果类别存在排序问题为定序型，若不存在排序问题则为定类型，非类别型是字符串型。

对于非数值型，我们需要进行类别转换，即将非数值型转换为数值型，以方便机器学习算法后续处理。

对于定序型，我们可以使用序号编码，如“refund\_rule“变量，可以按照：엄격：3、일반：2、유연：1的对应关系赋予一个数值ID，转换后依旧保留了级别关系。

对于日期变量，如：“4월 2013”，该变量实际意义为：2014年4月注册，为让该变量有实际意义，我们计算这个日期与2019年1月的差值，如4월 2013转化为：（2018-2013）\*12+（13-4）=69，实际意义为注册时间是69个月前。

## 2.3条。处理缺失值

这里的缺失值需要从两个角度出发去理解，首先是变量角度的缺失值，部分变量在统计过程中具有很多缺失值，我们认为这个问题的界限是百分之80，如果一个变量它的百分之八十都是缺失值，那么他在本身的处理过程或者与其他变量的综合分析过程中很可能表现不佳，因此是我们需要消除的。

其次是观测值的缺失值，也就是每一个airbnb的某些变量可能存在缺失，很显然在算法中这些缺失会导致这个观测值无法被结合，也会导致算法的综合性能下降。因此对于这种缺失值我们的策略是，如果某个观测值，在经过上一步处理后仍然有缺失值，我们将这个观测值删去。

## 更贴近业务的处理

在仔细观察提供的数据集后，我们发现结果上面两步处理的数据集存在一个问题。在数据集中排名靠后的变量，根据其实际意义来看是衡量该airbnb是否提供一些额外功能或设备的变量，但经过初步的方差分析，我们发现这样一个问题，部分变量中超过百分之90的观测值全为0，也就是说提供该服务或设备的airbnb极少，显然这样的变量不会对模型起到任何正面作用（主要是可能产生过拟合，或不满足统计学上的显著性要求），因此我们模仿第一步的做法，设置阈值为80%，也就是说超过百分之80的airbnb提供的服务或设备，该变量才可以参与综合分析。

## 变量观察

经过以上步骤的处理，进入分析步骤的变量已经被我们选中，下面我们观察数据的分布与关联。

## Modelling & Evaluation

## 4.1. 线性回归

### 4.1.1 模型介绍

多元线性回归模型通常用来描述变量y和x之间的随机线性关系，即：

式中，x,y分别为回归分析的自变量和因变量，是回归系数，是随机误差变量。

### 4.1.2 回归结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Pr(>|t|) |  | Estimate | Pr(>|t|) |
| number\_of\_bedrooms | 104977.2 | < 2e-16 | dryer | 14605.9 | 0.66068 |
| number\_of\_beds | 38049.9 | 2.38E-08 | necessities | 21872.7 | 0.36092 |
| number\_of\_bathrooms | 166792.2 | < 2e-16 | heating | 27223.9 | 0.37728 |
| number\_of\_individual\_bathrooms | 97187.1 | 7.42E-06 | refund\_rule | 25133.1 | 0.00224 |
| number\_of\_shared\_bathrooms | 28174.2 | 0.08555 | kitchen | 21596.2 | 0.2529 |
| fire\_extinguisher | 5238.2 | 0.66329 | hanger | -4404.2 | 0.8047 |
| number\_of\_language\_of\_host | 2154.9 | 0.61343 | hair\_dryer | 28837.0 | 0.08632 |
| response\_rate | -18055.2 | 0.7621 | shampoo | 16075.5 | 0.2698 |
| response\_time | -238.4 | 0.74512 | overall\_rating | 71864.2 | 3.92E-06 |
| registration\_date | -66.9 | 0.78808 | iron | -5707.2 | 0.70335 |
| number\_of\_reviews | -233.1 | 0.0243 | fire\_sensor | -40615.4 | 0.0274 |
| air\_conditioning | 28658.3 | 0.14753 | first\_aid\_kit | -13622.7 | 0.25908 |
| Wireless\_Internet | 4086.8 | 0.93758 | TV | 22721.1 | 0.08468 |
| washing\_machine | -7725.0 | 0.81678 | elevator | 39257.0 | 0.00198 |
| laptop\_work\_space | 18670.6 | 0.15657 | hot\_water | 16314.1 | 0.15229 |
| carbon\_monoxide\_sensor | 11306.5 | 0.44175 | cable\_TV | 21171.1 | 0.08573 |
| (Intercept) | -541662 | 1.70E-06 |  |  |  |

在回归的结果中，Estimate值可以认为是该自变量对应的系数，这个值的正负表示该变量对价格的正向或负向影响，即在示例数据的统计口径下，如果该项值为正数，那么这个变量与因变量之间存在正相关性，该项值为负数，这个变量与因变量之间存在负相关性。

Pr的理解涉及到到假设检验理论。在许多自变量共同影响着一个因变量的关系中，假设检验理论可以判断哪个（或哪些）自变量的影响是显著的，哪些自变量的影响是不显著的。那么在最终的模型中可以将影响显著的自变量加入模型中，而剔除影响不显著的变量，根据相关资料，一般认为pr值小于0.05表示该变量对应变量的影响是显著的。

综合上述内容，我们发现在对airbnb价格有显著影响的变量有：

number\_of\_bedrooms, number\_of\_beds, number\_of\_bathrooms, number\_of\_reviews, refund\_rule, elevator, overall\_rating, number\_of\_individual\_bathrooms,

显然，一些基本条件能显著影响airbnb的价格，但这些变量对价格的影响对房东来说没有任何参考价值，因为房东在一般条件下并不能改变这个房间的基本条件。因此，排除这些变量的影响，剩下的变量对价格的影响是我们所关注的。

refund\_rule

可以看到，退改规则与酒店价格的相关性十分强烈，这可能是由于合理的退改政策可能激发用户预定酒店的兴趣，并促成交易。

overall\_rating

综合评分能影响airbnb的价格，两者具有正相关性，这似乎也不难理解，很高的评分可能是由于较好的基础设施和服务。但这也给房东一些提示，虽然airbnb的基础条件短时间内无法提升，但通过对客户的服务和关怀，可以一定程度上提升评分，进而转化为价格的提升。

Elevator

关于电梯与价格的强烈正相关性，我们认为有一些奥秘存在。显然，airbnb的电梯给客人最直接的住宿感觉，如果一个电梯里味道太难闻，那么客人对酒店的印象分就会大打折扣，也会在心里留下酒店的卫生是否真的干净这样的疑问。可以看出，变量列表中有很多可选设备，如熨斗，hair\_dryer, shampoo, 但电梯的重要性名列前茅。

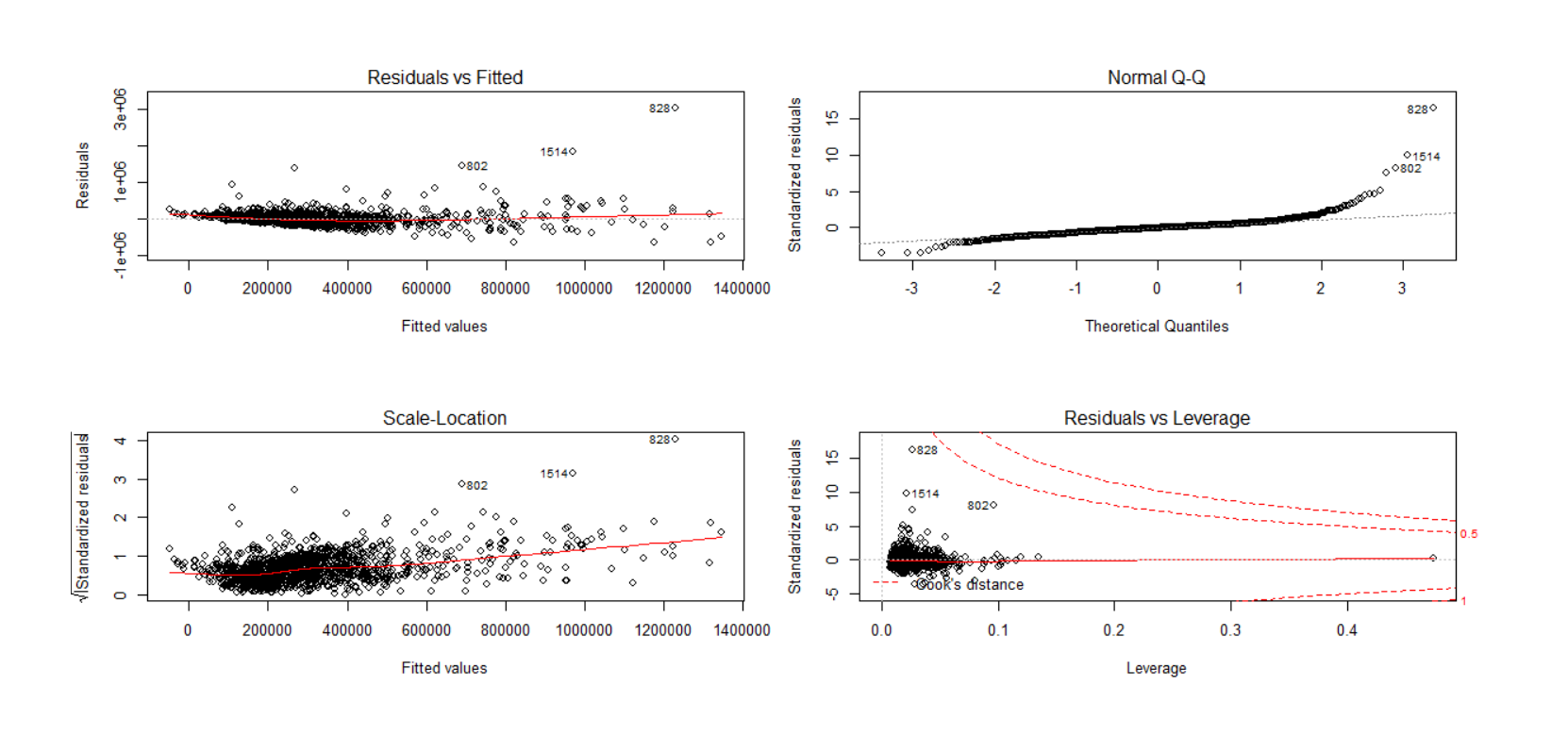
number\_of\_reviews

这是几个相关性较强的变量中唯一一个具有负相关性的变量，我们认为这可能是由于，一个价格过高的酒店入住者会较少，那么积累的review也不会多。这也给我们启示：对于部分入住者较少的高档酒店，尽量让每一位入住者都发表评论，是一个很好的展示高端airbnb的窗口，可能能大幅度提升酒店的曝光度，进而影响价格。

### 4.1.3 模型诊断

当自变量为分类变量、因变量为连续变量时，也是可以采用线性回归的。只是在更多的时候，这种类型的分析我们更关注的是组间差异，而不是线性回归预测，通常采用方差分析或者t检验，尤其是自变量只有1个的时候。模型假定不同的组来自同一个总体中的抽样，各组（严格说应该是各个变量）的残差服从同一个正态分布，不同组的残差均服从同一个均数为0标准差为σ^2的正态分布。在实际考察的时候我们往往直接考察固定的自变量值（不同的变量）对应的因变量值是否呈正态分布。

当然我们也可以采用了线性回归进行分析，为了消除赋值带来的误差，多分类的自变量在线性回归模型中需要设置成哑变量，结果同方差分析是一致的。对于上一步得到的回归模型，我们做了回归诊断如下：



正态分布假设

连续性变量取值往往较多，即使我们的样本量足够大，自变量的每一个固定值有多个取值，这种考察正态性的工作量也会变的很大。这种情况下，对正态性的考察回归到对所有残差的考察反而更简便。基于这一假设，数据的有效性得到了保证

离群值

因为过大或过小的数据可能会影响到分析结果，尤其是在做回归的时候，我们需要对那些离群值进行处理。在进行基本处理时，我们运用了标准化方法，用于多个不同量级指标之间需要互相比较的时候。从图中可以发现，整体数据中离群值较少

线性模型合理性

分位数回归则可以理解成加权的最小二乘估计，一般的最小二乘目标是估计参数的期望（均值），而分位数回归则是直接估计出分位数的值。分位数回归提出的原因，就是因为不希望仅仅是研究参数的期望，而是希望能探索参数的完整分布状况，或者说可能在某些情况下我们更希望了解参数的某个分位数。观察QQ图可知，在主要区域内，分位数期望服从总体，回归模型比较合理。

残差分析

## 4.2，决策树

### 模型介绍

决策树是一种基本的分类与回归方法，它是一种描述对实例进行分类的树形结构。在使用分类树进行分类时，从根结点开始，对实例的某一特征进行测试，根据测试结果，将实例分配到其子结点。这时，每一个子结点对应着该特征的一个取值。如此递归地对实例进行测试并分配，直至达到叶结点。最后将实例分到叶结点的类中。

### 特征工程

在数据集中，因变量作为价格，是一种连续属性。根据决策树的划分条件，是选取一个最合适的特征属性，然后将集合按照这个特征属性的不同值划分为多个子集合，并且不断的重复这种操作的过程。出现了连续属性显然会影响这一过程。但是我们又想让价格这个特征属性参与到决策树的建立中，我们对其进行处理，用来满足决策树的需要

例如，在直角坐标系中x轴上有无数个点，这些点就像对应的密度值一样，我们无法计算出其总数，所以不能用这个点进行划分。但是直角坐标系中还有y轴，在y轴左边的点x值都是负的，在y轴右边的点都是正的，所以可以说y轴将x轴划分为正负两个部分。那么我们就可以利用这种划分方法将密度值进行划分多个子区域，这样密度值的子区域是有限的，所以就可以进行计算了。

根据这一理论，我们对价格这一变量进行划分，我们计算了价格变量的四分位数，将整体区间分为四个部分，并转化为分类变量，从而进行进一步的运算。

### 模型主要结果

|  |  |
| --- | --- |
| Variable | importance |
| number\_of\_bedrooms | 40 |
| number\_of\_bathrooms | 25 |
| number\_of\_beds | 19 |
| registration\_date | 4 |
| number\_of\_shared\_bathrooms | 3 |
| washing\_machine | 3 |
| dryer | 2 |
| elevator | 1 |
| number\_of\_reviews | 1 |
| number\_of\_individual\_bathrooms | 1 |
| hot\_water | 1 |

## 4.3综合诊断

4.3.1理论介绍

ROC曲线

ROC曲线是根据一系列不同的二分类方式（分界值或决定阈），以真阳性率（灵敏度）为纵坐标，假阳性率（1-特异度）为横坐标绘制的曲线。传统的诊断试验评价方法有一个共同的特点，必须将试验结果分为两类，再进行统计分析。ROC曲线的评价方法与传统的评价方法不同，无须此限制，而是根据实际情况，允许有中间状态，可以把试验结果划分为多个有序分类，如正常、大致正常、可疑、大致异常和异常五个等级再进行统计分析。因此，ROC曲线评价方法适用的范围更为广泛。

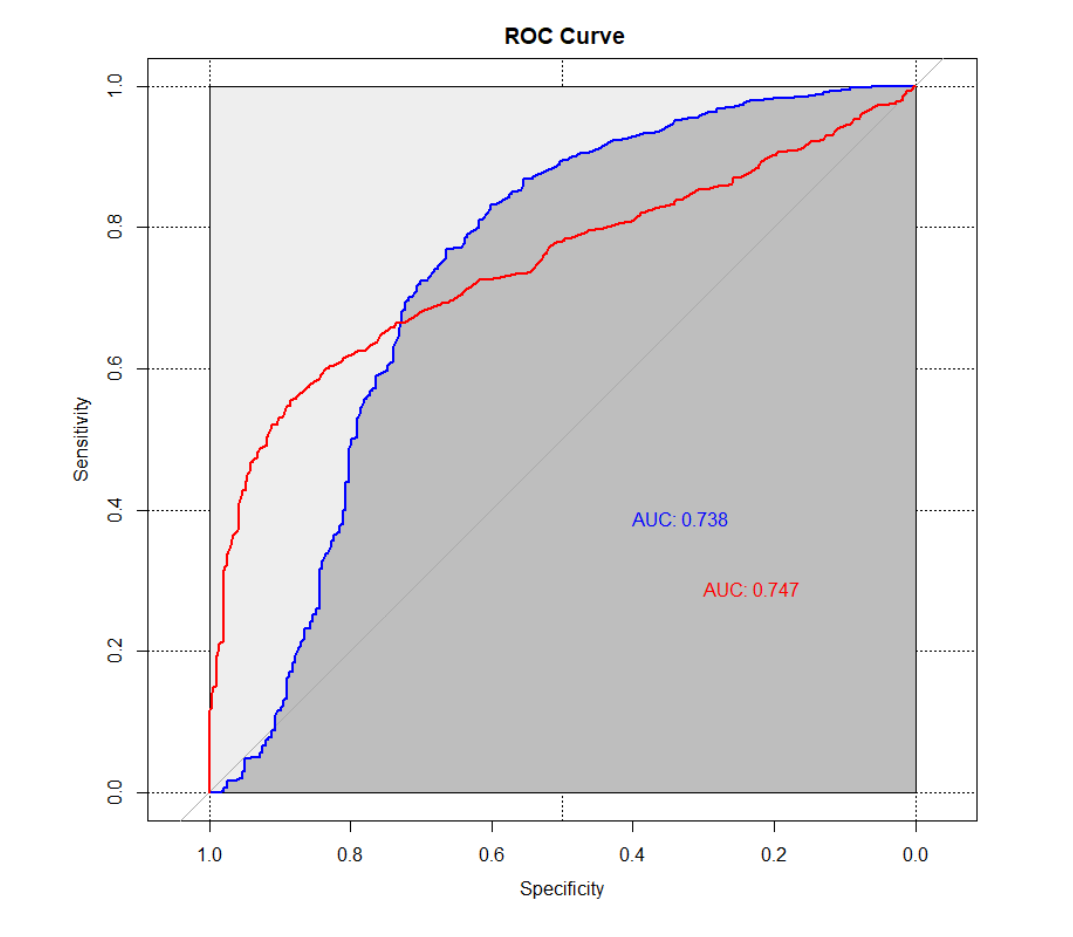
AUC值

AUC是一个类似于二元模型的一致性或c统计量的度量。这是具有正类的观察将具有比在负类中的观察更大的预测概率的概率。如果AUC = 1，则表示模型有完美的预测。如果AUC = 0.5，则意味着模型无法区分类。它在逻辑回归中的行为与r^2相似，因为添加更多预测因子会增加AUC。因此，在分析中包括对外部数据的交叉验证或验证非常重要。

我们建立模型的目标是的高更高的AUC

4.3.2对比诊断

两个模型的ROC曲线如下，其中蓝色曲线表示线性回归算法，红色曲线表示决策树算法



可以看出，两种算法的AUC值位于0.73-0.75范围内，说明两种算法性能相近，总体来说表现良好。

聚类分析

聚类分析是根据在数据中发现的描述对象及其关系的信息，将数据对象分组。目的是，组内的对象相互之间是相似的（相关的），而不同组中的对象是不同的（不相关的）。组内相似性越大，组间差距越大，说明聚类效果越好。

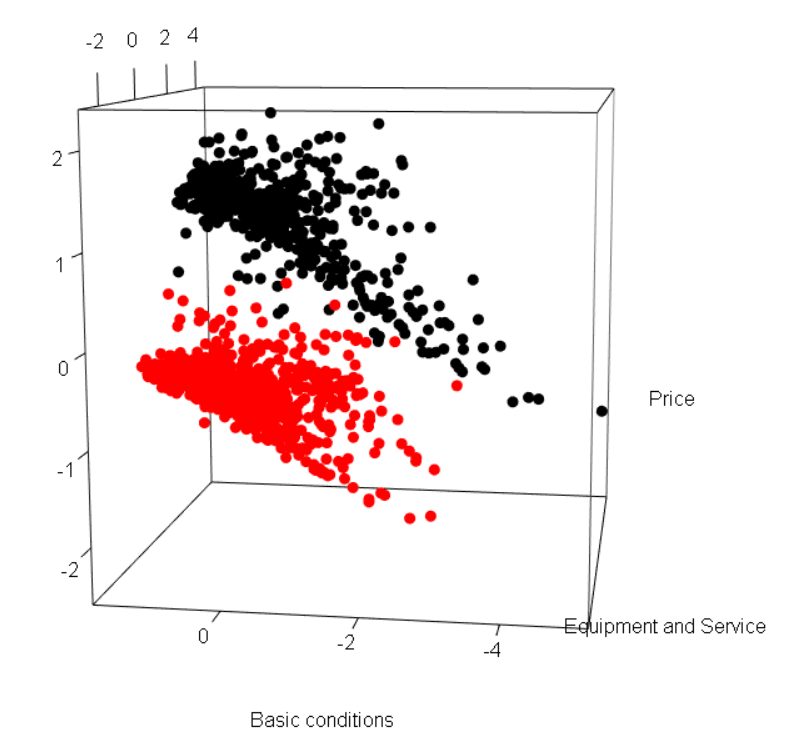
聚类分析可以对变量进行自然分组，从而发现自然类别，提取出变量特性，使我们对观测值间的关系有进一步的理解。

主成分分析

主成分分析是一种多变量统计方法，它是最常用的降维方法之一，通过正交变换将一组可能存在相关性的变量数据转换为一组线性不相关的变量，转换后的变量被称为主成分。主成分分析可以起到降低维度的作用，对当前数据集（拥有三十多个变量）非常适合

提取主成分，利用不同的主成分拟合不同的观测值，最终得到3个观测值序列，选取每个特征值序列中占比最高的因素，作为该序列的代表因素。

最终将这三个特征值序列作为聚类的原料，进行聚类分析。



根据我们的选取规则，三个主成分的代表因素分别为“Equipment and Service”，“Basic conditions”，“Price”，根据途中的聚类结果，我们发现不同airbnb根据价格因素显著的分为两部分，这两部分中的Basic conditions 和 Equipment and Service可能具有相同的水平，但价格差距却较大，这也佐证了我们前面的观点：一些软实力能显著影响airbnb的价格

# 4. Insights and Implications

## 4.1. Evaluate Business Problem

Problem1:什么因素影响airbnb价格

Problem2:什么因素可以被房东改善，从而提升airbnb价格

## 4.2. Recommendations & Suggestion

( 一 )根据设施水平调整价格

基本条件相同的airbnb的经营情况不一定相同，受较多附加设施影响，airbnb需根据不同附加设施制定不同的目标。在附加设施比较缺乏的airbnb，可以降低目标利润以降低房价并提高入住率。相反，在基本条件有限但附加设施比较充足的房型，可以增加目标利润，提高房价，增加房间的营业收入。虽然低附加设施房间可用于支付固定成本的收入较低，但对抵消经营活动的固定成本会产生积极影响；在高附加设施房间，这种抵消会增加。

( 二 ) 基本条件能决定价格

airbnb的运营不能根据临时营业额来提高价格。有必要根据airbnb的基本条件（如房间类型，房间面积）制定合理的价格水平。这样不仅可以保证综合的利润率，还可以在一定程度上保持良好airbnb品牌的形象。让airbnb价格满足客户的基本预期。

( 三 ) 根据服务水平区分定价

airbnb的定价基于airbnb自有品牌，商业目的，季节和客流。对于服务水平较高的高端airbnb，价格在一定范围内提高，符合airbnb品牌和当地airbnb价格，在客户入住体验良好的前提下提升airbnb利润。当服务水平较低时，也不能盲目打折。折扣应该与airbnb的基本条件和附加设施水平相适应，保证airbnb自身的利益。

4.见解和启示

4.1。 评估业务问题

4.2。 建议与建议

1、保持优势区

保持优势区属于决定airbnb价格水平的因素。位于该区域的因素主要包括房间数量，卫生间数量，房间面积，床数量等基本指标。说明这些因素是旅行者在选择airbnb时非常重视的因素，同时也是实际住宿体验过程中感觉比较满意的因素。因此，airbnb应继续维持这些项目的服务水准，并且在日后重视这些项目的投入。

2、集中改善区

集中改善区是能进一步影响airbnb价格水平的因素。位于该区域的因素有各种airbnb提供的附加设施，如烘干机，洗衣机，吸尘器，电视机，香波等。这些因素的重要性得分高于平均水平，说明旅行者较为重视这些因素，airbnb房东应该正视顾客旅行者的需求，并采取有效措施来改善不足之处。

3、最后补充区

次优先改善区是最后能提升airbnb价格水平的因素。主要包括回复时间，退改规则等。airbnb房东在基本条件和附加设施都无法提升时，可以考虑对这些因素进行改善。

。

内容提要业务问题作为一组数据分析人员，我们要解决的业务问题是“一部成功电影的关键属性是什么？”根据我们的业务问题，我们将使用数据挖掘技术进行研究和分析，以提供清晰而有说服力的见解。我们的团队旨在产生一个模型，以使合适的决策者能够使用它来实现高投资回报率（ROI）。

业务问题

针对这份airbnb数据，我们要解决的业务问题是“如何让airbnb房东获得更高的利润”，根据我们的业务问题，我们将分析什么因素影响airbnb的价格，结合现实场景提出如何让房东把airbnb卖出更好的价格。

数据挖掘技术

可以帮我们解决业务问题的数据挖掘技术是回归分析，决策树分析和聚类分析。我们先使用回归分析，分析被选中的变量与airbnb价格之间的关系，对影响的方向和显著性做了研究。其次，我们使用决策树方法，挑选出对airbnb价格影响最大的部分变量，并利用模型将这种影响量化。最后，我们采用聚类方法根据airbnb的价格，设施条件和服务水平，将airbnb分为不同的类别，探究不同类别之间的差异。

模型结果

回归分析，决策树分析和聚类分析得到类似的结果，决定airbnb价格的最大因素是airbnb的基本条件，也就是房间的类型，面积等；其次，房间附带的设备如洗衣机，烘干机，电梯等也能影响airbnb的价格，最后，airbnb的服务水平也对它的价格有一定影响

见解和建议，

对每个数据分析模型进行分析后，我们就如何提高airbnb房东的收入和利润提供了三个方向：首先，着力于改善基本条件，其次，提供尽可能多的设施服务，最后，提供更好的服务水平。

数据挖掘技术可以帮助我们解决业务问题的数据挖掘技术是文本挖掘，聚类和线性回归。我们报告中的第一个任务是文本挖掘，该文本挖掘用于识别跨不同电影通常使用的流行关键字。随后，第二任务包括使用不同的聚类方法来根据识别出的流行关键字对电影进行分组。最后，我们将通过分析每部电影的投资回报率与其他关键变量之间的关系来执行线性回归。在我们的回归分析中，使用了两种不同的模型来预测电影的投资回报率。这是因为使用的部分变量仅在Facebook成立于2008年之后才相关且适用。

模型结果在文本挖掘和聚类之后，“女性”一词出现得最多。 21个最常用的术语分为5类：成人，犯罪，戏剧，都市和青年。扩展聚类分析的结果后，流行关键字和通用关键字并不等同于ROI的增加。相反，对提高ROI具有最重要意义的关键预测指标是预算较低，IMDB得分较高和持续时间较短。

见解和建议对每个数据挖掘任务的输出进行分析之后，该报告就如何解决当前的业务问题向决策者提供了建议和建议。一些建议包括：制片人专注于不围绕陈词滥调主题的内容，降低预算，无视导演的生产力以及缩短电影播放时间