北京市短租房源推荐

小组成员：徐一丹 3120191063
刘嘉 3220190842
刘玉洁 3220190848

一、 摘要

北京市短租房源推荐旨在分析北京市短租房源情况，并对各房源进行评论挖掘，通过 LSTM 对评论数据进行文本情感分析，最后根据 LSTM 分析结果及租房者基本需求进行短租房源推荐，以帮助租房者快速找到满意的房源。

二、 背景

线上短租房市场发展势头迅猛， 共享经济为闲置的房屋资源提供了合理利用的途径， 避免了资金和空间的闲置浪费。当互联网的“订单经济”与旅游住宿相互交叉， 闲置房屋的短期出租交易作为共享经济在旅游届产生的新兴产业，由于其性价比高、租房周期随意、租房时间机动性强，正在逐渐代替传统酒店、旅馆住宿。当前，中国短租房交易较为活跃的地点主要集中在一线城市和东部城市，据统计，这些地区的房源供给和用户占比均在 60%以上。可以预见，短租房未来市场潜力巨大，房源数、用户数、交易额等都将持续增加，而短租房交易的发展也将带动保洁、装修、维修、保险等相关行业的发展。同时,随着线上短租房交易平台巨头 Airbnb 的强势进驻,中国短租房交易市场迎来新一轮的机遇和挑战。促进共享经济中的短租房业务可持续发展需深入量化分析租房价格与数量的分布。

本项目选择北京市这一短租房交易活跃的城市作为分析城市。首先对北京市的短租房源数据集进行数据预处理和基本的探索性分析，从而挖掘当前北京市短租房源市场情况。然后利用短租数据集中的房源评论信息进行 LSTM 文本情感分析，作为房源推荐关键信息。最后根据租房者实际需求，在符合条件的房源中，依据 LSTM 文本情感分析结果进行房源推荐，为租房者推荐出满足条件的十个最佳房源以供参考。

三、 数据集

本项目共使用了两个数据集：

- **天池短租数据集**

本项目使用天池短租数据集挖掘短租房源相关信息，该数据集来自天池大数据竞赛，具体包括如下信息：

1、短租房源基础信息：房源、房东、位置、类型、价格、评论数量和可租时间等；

2、短租房源时间表信息：房源、时间、是否可租、租金和可租天数等；

3、短租房源的评论信息：房源 id、评论日期、评论内容、作者信息等；

4、北京的行政区域划分信息。

- **图书数据集**

由于天池短租数据集没有提供评论的情感倾向标签，我们使用了一个有标签的图书评论数据集来进行 LSTM 模型的训练，然后使用训练好的模型对天池短租数据集的评论的情感倾向进行预测。

四、 数据探索性分析

本部分对北京市短租房源数据集进行预处理和基本的探索性分析，并可视化相应结果，并给出结果分析说明。

1. 数据缺失值处理

首先查看数据集的基本情况：

```
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              28452 non-null  int64
 1   name                            28451 non-null  object
 2   host_id                         28452 non-null  int64
 3   host_name                       28452 non-null  object
 4   neighbourhood_group             0 non-null      float64
 5   neighbourhood                   28452 non-null  object
 6   latitude                        28452 non-null  float64
 7   longitude                       28452 non-null  float64
 8   room_type                       28452 non-null  object
 9   price                           28452 non-null  int64
 10  minimum_nights                  28452 non-null  int64
 11  number_of_reviews               28452 non-null  int64
 12  last_review                     17294 non-null  object
 13  reviews_per_month               17294 non-null  float64
 14  calculated_host_listings_count  28452 non-null  int64
 15  availability_365                28452 non-null  int64
dtypes: float64(4), int64(7), object(5)
memory usage: 3.5+ MB
```

可以看到，数据集共有 16 个属性，28452 条数据，其中 name、neighbourhood_group、last_review 和 reviews_per_month 等 4 个属性存在数据缺失。我们根据这几个属性的实际情况进行不同的缺失值处理，其中 name 属性只有一条数据存在缺失情况，直接删除该条数据即可，neighbourhood_group 所有属性均为空值，直接将该属性列删除，last_review 和 reviews_per_month 数据缺失量较大，而且无法根据其他属性列进行填充，如果删掉相关数据会影响短租数据分析的结果，而且也不合情理，因此我们对此不做处理。

2. 数据异常值处理

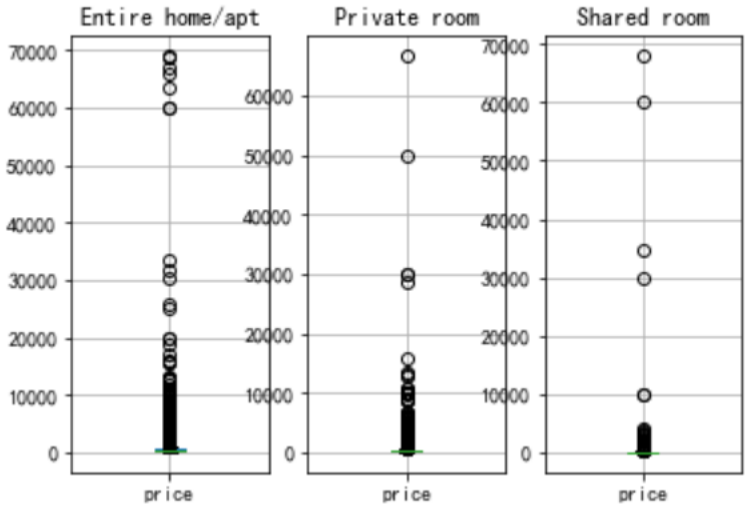
接下来我们查看各数值属性的统计情况：

	price	latitude	longitude	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	28452.000000	28452.000000	28452.000000	28452.000000	28452.000000	17294.000000	28452.000000	28452.000000
mean	611.203325	39.983225	116.442000	2.729685	7.103156	1.319757	12.818290	220.342120
std	1623.535077	0.186984	0.204796	17.920932	16.815067	1.581243	29.261321	138.430677
min	0.000000	39.455810	115.473390	1.000000	0.000000	0.010000	1.000000	0.000000
25%	235.000000	39.897330	116.355283	1.000000	0.000000	0.290000	2.000000	87.000000
50%	389.000000	39.930905	116.434665	1.000000	1.000000	0.800000	5.000000	209.000000
75%	577.000000	39.990470	116.491122	1.000000	6.000000	1.750000	11.000000	361.000000
max	68983.000000	40.949660	117.495270	1125.000000	322.000000	20.000000	222.000000	365.000000

通过对各数值属性的基本分析，房源所在经、纬度、评论数、每月评论数、可出租房屋数和可租天数等属性的情况基本符合常理，但价格和最短居住日期可能存在异常值，我们绘制相应盒图并分析。

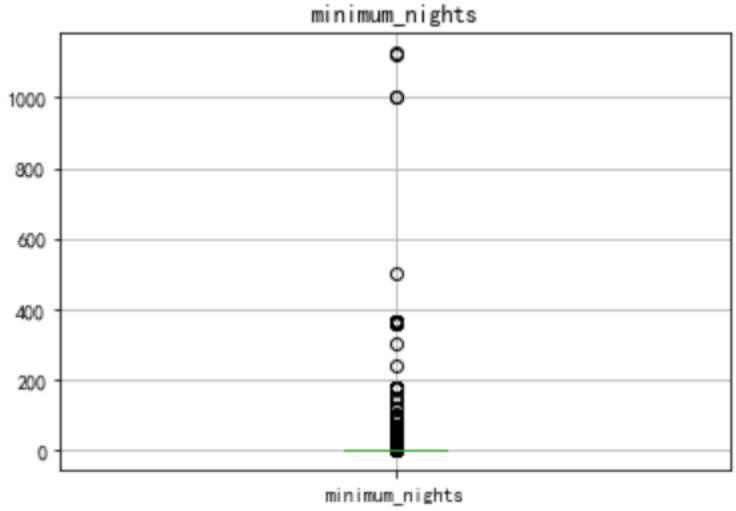
价格：通过分析最大最小值情况，可以发现房源最小值为 0，显然不合常理，考虑删除价格为 0 的房源。而房源最大值为 68983，虽然北京可能存在四合院或

别墅等价格较贵房源，但这个价格依然较高，根据分析房源存在套房、独立房间和合租房三种类型，不同类型的房源价格可能差异较大的，因此我们根据房源类型绘制盒图。



根据盒图，可以看到，套房的均价最高，合租房的均价最低，符合我们通常的认知，但是三种类型的房源最高价都达到接近 70000，这显然是不合常理的，因此，我们需要对异常值进行处理。根据盒图分布情况，删除价格在 50000 以上的整租房源，价格在 20000 以上的独立房源，以及价格在 10000 以上的合租房源。

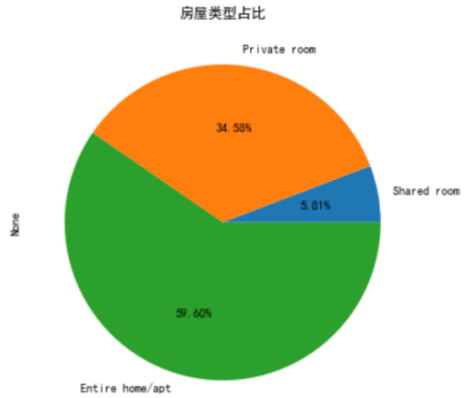
最短居住日期：最短居住日期大部分为 1 天，符合基本情况，但最大值为 1125，接近 3 年，不大符合短租数据集情况，绘制盒图

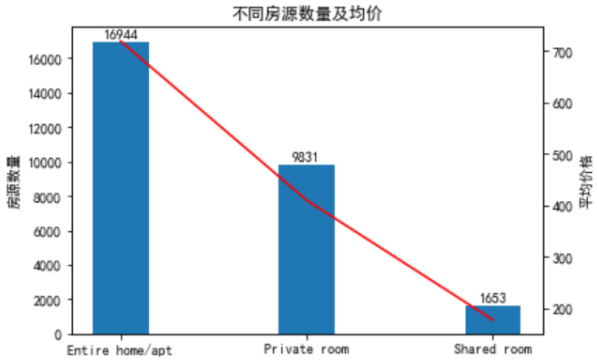


将最短居住日期在 600 天以上的数据删除。

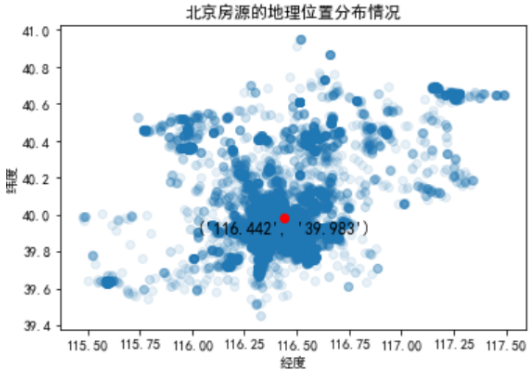
3. 房源信息及价格分析

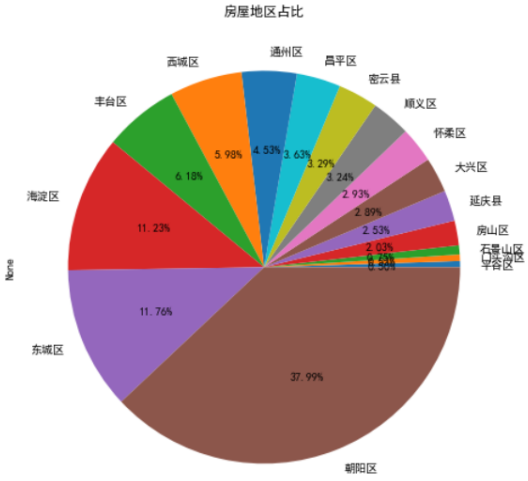
北京市短租房源分为套房、独立房间和合租房三种，如下图,其中套房数量最多,均价也最贵，大概在 700 左右，而合租房数量最少，均价最低，大概在 150 左右，整体来看，北京市短租房源价格偏高





下面分析房源位置分布情况，根据经纬度绘制房源位置分布图，以及地区占比图，如下图，可以发现北京市的短租房源主要集中于市区，以朝阳区的短租房源密度最大，而郊区的房源数量较少。



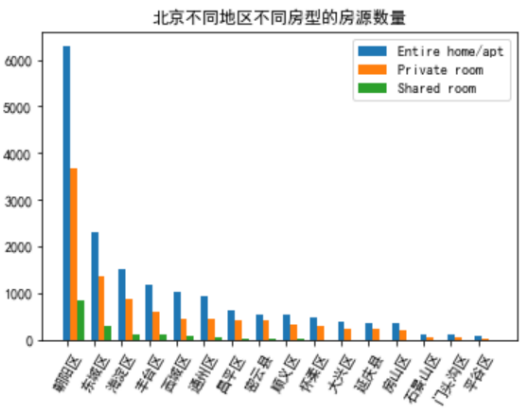


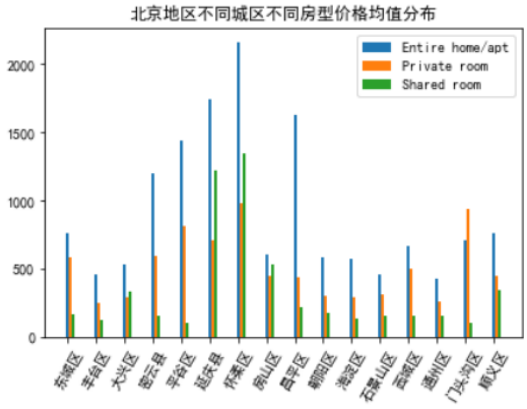
接下来分析北京市房源价格与位置关系。下面左图是北京是不同区不同房源的数量图，右图是不同地区不同房源的价格图，可以看出：

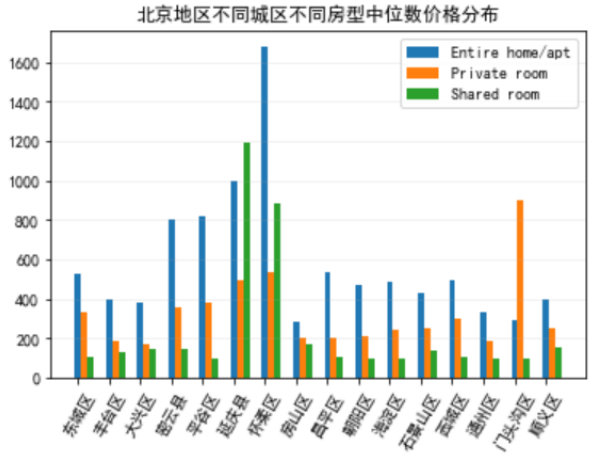
（1） 北京市的合租房主要集中在市区内，周围几个区县内合租房数量较少；

（2） 在市区内合租房源的价格差异不大；对于套房，东城西城均价相对较高，而中位数与其他区接近，考虑这两个区存在价格较高的四合院，因此拉高了均价；对于独立房源，东城西城这两个区价格相对更高；

（3） 而郊区的房源价格普遍高于市区内，可能是由于郊区房源主要位于景区内度假休闲娱乐，因此价格相对较贵，其中门头沟区的独立房间均价与中位数均高于套房，怀柔延庆两区合租房均价中位数高于独立房源，可能是之前进行数据预处理时比较粗暴，漏掉了一些异常值。







4. 挖掘租房者核心需求

根据评论制作词云，查看评论关键词,根据词云信息可以看出房间干净是评论提及最多的，此外交通方便和房东热情的提及率也非常高，可以看出，房间干净整洁是租房者的核心需求，在此基础之上，交通方便与房东热情也是租房者较为关注的。

五、 LSTM 文本情感分析

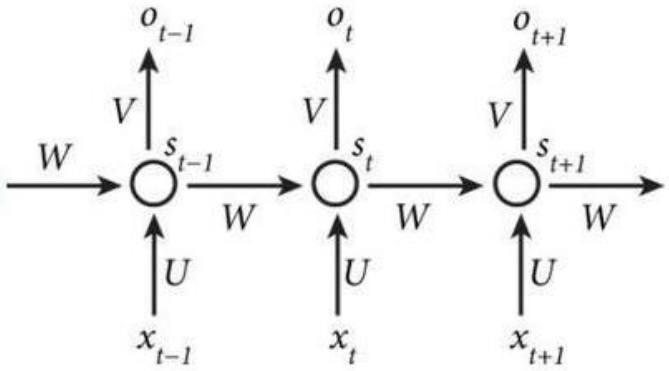
1. 问题的提出

本文为了本文所使用的短租数据集只有评论信息，并没有评分等易于量化和排序的信息，这样的话在做房源推荐时若以 id 或名称进行排序，会让消费者花时间用于排除糟糕的房源，所以需要对评论文本做文本情感分析。

文本情感分析又称倾向性分析，是指对一段文本所包含的感情色彩进行分析和推理的过程。文本情感分析的方法主要有基于规则的方法和机器学习的方法。近年随着机器学习的热度上升，各种使用神经网络的情感分析方法横空出世。神经网络中被广泛用于文本处理的结构是循环神经网络（RNN）。而 RNN 具有梯度消失，训练过慢等问题，在长序列中尤为明显。作为 RNN 的一种改进体，长短期记忆网络（LSTM）具有长期记忆，训练方便等优点，应用越来越广。本文使用基于 LSTM 的情感分析模型来对评论信息进行处理。

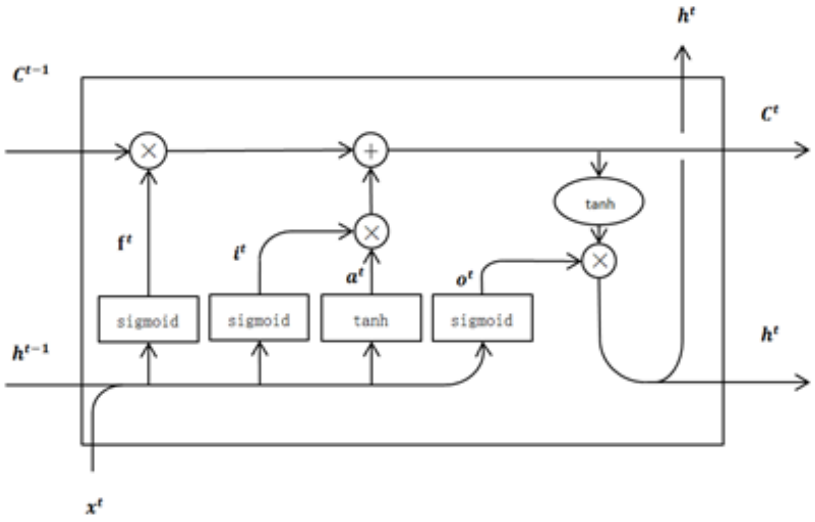
2. LSTM 模型结构

RNN 的基本结构图如下：



可以看到比起最简单的神经网络，RNN 的特点在于除了推理方向上的信息流动，还有文本顺序方向的信息流动。这样的好处是网络在对某一个词进行预测的时候，不仅考虑词本身的信息，还考虑了之前输入的词的信息。例如，一个形容词本身的含义是积极的，正面的，但如果前面有了否定意义的词，意思就会相反。所以比起简单的神经网络，RNN 在处理文本问题更有优势。

但是，由于 RNN 的网络深度有层数和循环次数（也就是序列长度）两方面，RNN 在处理长序列时问题网络加深，更易产生深度神经网络的梯度消失问题，导致训练进程缓慢。所以 RNN 的一种变体 LSTM 的应用更多。LSTM 的单元结构图如下：



由于具有门控的思想，LSTM 可以选择是否要记忆更多的内容，从而具有遗忘不重要信息的能力，使得梯度消失的问题得到缓解。

3. 模型的构建

本文模型使用 keras 包在 python3 进行搭建和训练。

在进行 LSTM 训练之前，数据先进行分词。因为若用字为单位则会有很多误解的情况发生,同时输入长度也很长,影响模型效率和效果。中文词语边界模糊，中文分词是文本领域另一个重要问题。本文使用结巴分词进行分词。

有了词表，我们构建字典，即词到索引和嵌入向量的映射。然后训练 word2vec 模型。这一步的目的是将词汇初步编码为表示更高效，更有解释性的词嵌入向量，包含更多的可解释的语义信息，为后续处理降低难度。

模型的框架是层级结构，各层分别为：

第一层：嵌入层。用于把 one-hot 向量变为词嵌入向量。

第二层：hidden size=50 的 LSTM 层。用于编码词信息。

第三层：全连接层。用于把 50 维的隐向量变换为三维的向量，代表在三种类型上的输出值。

第四层：softmax 层。用于将三种类型上的数值转变为概率。

除了模型结构，模型的其他信息有：优化器为 adam，损失函数为交叉熵，batch_size 为 32。

4. 模型的训练和推理

模型在图书数据集进行训练，在短租数据集进行推理。

图书数据集分为好评、中评和差评三部分，分别设定标签为 1,0，-1。送入模型进行训练。训练结束后保存模型。

在推理时，加载训练好的模型参数，预测。将预测结果为 1 的概率保存，之后房源推荐使用这个信息。

六、 短租房源推荐

本部分依据上一部分所得出的 LSTM 文本情感分析结果作为推荐的关键指标，结合用户限定的基本条件，为用户推荐满足条件的十大最佳房源。

1. 计算各房源评分

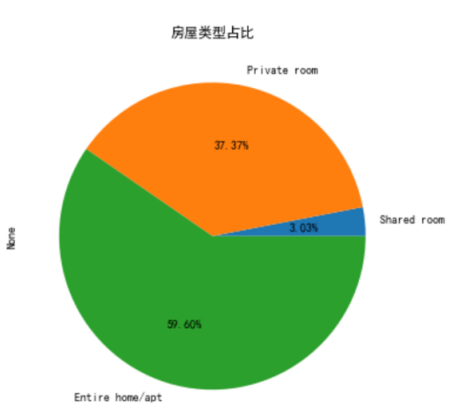
依据前一部分 LSTM 文本情感分析模型，求出每条评论的正向情感评分，属性值 positive_prob 为最终结果。观察 listing_id 列的数据可以发现，相同房源由于评论不同，所得出的结果也不相同，所以要对相同房源的评估评分取平均值作为房源的整体评分，并按照降序排列将其保存至 result.csv 文件中。

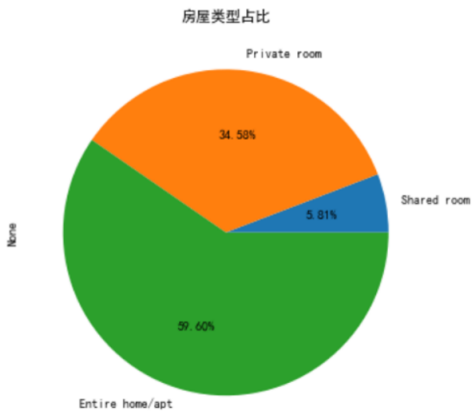
listing_id	id	date	reviewer_ic	reviewer_n	comments	positive_prob
44054	84748	2010/8/25	207019	Jarrod	Sev was very helpful. Sev showed us where to s	0.958531
44054	118384	2010/10/13	218723	Kimberly	We arrived in Beijing very early in the morning due to a delayed flight and Sev/East Apartments was very accommodating with helping us locate the building and getting us settled into the apartment. We were traveling in a group of 5 and found this apartment to be perfect for us- we stayed in a 3 bedroom apartment and there was plenty of space as well as a kitchen, two bathrooms and a washer (although, the washer did dye one of my white shirts a faint green)	0.919216
44054	436978	2011/8/11	609177	Emma	It is a really massive apartment and really comfortable. Fully equipped for all you need in Beijing. Clean, nicely furnished and great size. Sev is an excellent host, we couldn't have survived without him,. Always there to lend his advice and help. It was so helpful! We really appreciated him. Thank you :)	0.994668
44054	1118657	2012/4/12	1787536	Andreyna	Sev was incredibly helpful, showed us around the neighborhood and was available in case any questions or concerns arose. Not only that but they also offered us a tour service to the great wall which was really good! The apartment was well located with near	0.99146

2. 数据可视化

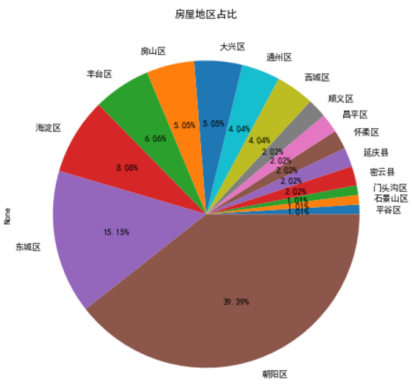
取前 100 位优质房源对数据进行可视化。

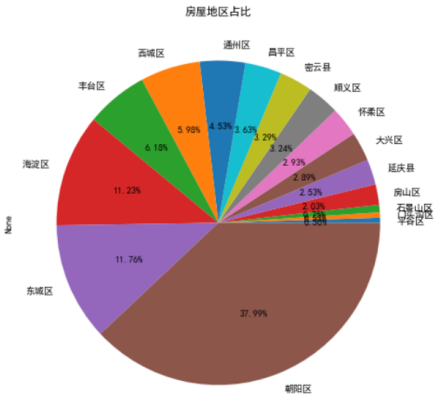
首先对房源类型占比进行可视化。通过优质房源类型占比（左）与所有房源类型占比（右）的对比，可以看出两者之间的差距并不大，尤其是在整租类型中，两者占比相同。但合租类型中，优质房源更多，但相差不大。



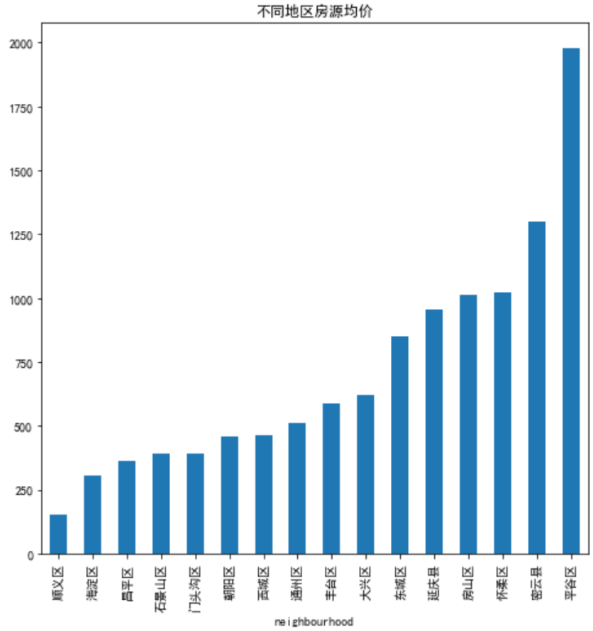


其次是房源地区占比。通过优质房源地区占比（左）与所有房源地区占比（右）的对比，可以看出东城区与海淀区两个地区的变化较大。东城区房源占比为 11%，但优质房源占比为 15%，增幅为 4%。可以看出东城区的房源中优质房源占比较大。而海淀区在房源占比与东城区持平的情况下，优质房源占比只有 8%，在海淀区租房时需要认真参考评论评分。





最后是关于房源价格。通过对不同地区优质房源均价的直方图可以看出，顺义区的优质房源均价很低，不到 200 元，而平谷区的优质房源均价达到 2000 元，相差了十倍左右。可见在北京不同地区的房价相差很大。建议用户在租房时，优先挑选顺义区、海淀区、昌平区等地的房源。



优质房源只挑选了前 100 位，数据量较小可能存在一定的误差，在挑选房源时还需根据实际情况认真考察。

3. 房源推荐

首先获取用户需求。我们将房源位置、类型以及价格区间作为索引条件，根据用户的请求进行第一步筛选工作。缩小房源选择范围。

```
#接收用户筛选请求
neighbourhood = input("请输入需要选择的房源位置（例如 东城区）：\n")
room_type = input("请输入租房类型（Entire home/apt，Private room，Shared room）：\n")
print("请输入您可以接受的价格区间")
price_min = input("请输入房价的下限：\n")
price_max =input("请输入房价的上限：\n")
```

```
请输入需要选择的房源位置（例如 东城区）：
朝阳区
请输入租房类型（Entire home/apt，Private room，Shared room）：
Private room
请输入您可以接受的价格区间
请输入房价的下限：
50
请输入房价的上限：
500
```

在满足用户条件的基础上，向用户推荐房源评分最高的十个房源。这些优质房源是根据过去用户的评价筛选出来的，具有很高的参考价值。为了使结果简单明了，我们只向用户展示了房源 id、简介以及月租价格，为用户节省时间，且目标性更强。

```
}]: print("向您推荐评价最好的十个房源：\n")
columns=['id','name','price']
for i in range(0,28451):
    for j in range(0,9):
        if listing.loc[i,'id']==m[j]:
            print(listing.iloc[i][columns].values)
```

```
向您推荐评价最好的十个房源：

[554123 'High-floor downtown studio #5' 416.0]
[1285333 'The closest  Wangjing 798 apartment' 255.0]
[1325945 '798 Fresh style Bedroom with Study ' 295.0]
[1484562 'Near Beijing subway an apartment' 127.0]
[1521874 'Nature style 1 bedroom near CHAOYANG Park' 168.0]
[1635589 '望京宝星园出租一间卧室(限女生)' 201.0]
[1779178 'big room near subway@ east chaoyang' 127.0]
[1834479 'small room in chaoyang, shuangqiao' 121.0]
[1834587 'welcome to beijing ,we have a room' 121.0]
```

七、 总结

通过本次大作业，我们小组成员进一步加深对数据挖掘课堂上所学到的数据处理知识以及数据挖掘算法的应用与理解，通过对短租数据集的探索性分析，分析了当前短租市场的情况并可视化相应结果，然后运用 LSTM 文本情感分析进行评论挖掘，最后利用分析结果结合租房者实际需求进行房源推荐。

本文的不足之处：

情感分析模型的训练数据集是中文的图书评论数据集，在短租领域的准确率会下降。如在测试时，会给“地方不好找”、“停车位太少了”这样的负面评价高分，而“这个价格，能有这样的房子太不可思议了”低分。另外 Airbnb 作为海外电商，北京作为国际化大都市，这份数据集有许多除中文外的其他文字，在这些评论中模型出错的概率会更高。

改进的办法：使用更合适的多语言（中文为主）的酒店等近似领域的带有标签的评论数据集作为训练集。