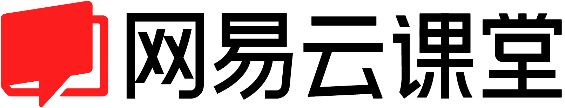
《数据挖掘》

短租房分析项目-实验手册



目录

[1 实验介绍 3](#_Toc29403)

[1.1 实验背景 3](#_Toc24687)

[1.2 实验目的 3](#_Toc14407)

[1.3 实验清单 4](#_Toc25332)

[1.4 实验环境介绍 4](#_Toc20611)

[1.5 前置理论知识 4](#_Toc29228)

[1.5.1 可视化图形介绍 4](#_Toc4393)

[2 短房租分析实验 6](#_Toc8157)

[2.1 实验总体设计 6](#_Toc29974)

[2.2 实验详细设计 6](#_Toc31739)

[2.2.1 材料准备 6](#_Toc21711)

[2.2.2 短房租价格特征分析 7](#_Toc15926)

[2.2.2.1 实验内容 7](#_Toc12552)

[2.2.2.2 实验小结 11](#_Toc27502)

[2.2.3 短房租价格影响因素分析 11](#_Toc25565)

[2.2.3.1 实验内容 11](#_Toc8987)

[2.2.3.2 实验小结 21](#_Toc19819)

[3 实验总结 22](#_Toc12622)

[3.1 实验总结 22](#_Toc19550)

# 实验介绍

## 实验背景

本案例中的数据来自于爱彼迎（Airbnb）网站2018-2019年度多伦多市的真实数据。数据集中包含listings数据集，约2万条数据，记录着所有的房屋信息，包括价格在内的及时项信息字段。另一个数据集名为calender，包含约650万条的租房交易数据，记录每一天每一所住房的入住信息。

通过对房屋价格按天/周/月进行汇总，我们能够从不同时间尺度上完成房屋价格信息的总结，查看房屋价格走势与波动规律。在分析走势的基础上，尝试分析房屋价格与房屋内部信息之间的相关关系和影响程度，能让我们更好完成对房屋价格的预测，提高对短房租数据分析的掌握程度。

本章实验难度：中级。

## 实验目的

本案例将掌握如何使用数据可视方法完成短房租价格走势、房屋价格影响因素的分析。

通过本实验的学习，您将能够：

掌握如何使用isnull.sum()、nunique()等函数完成数据预处理与数据清洗等操作；

掌握使用直方图查看房屋好评率数据、房价数据的分布情况；

掌握使用箱线图完成高级房、普通房价格的描述性分析与对比；

掌握使用条形图完成房屋多种类型的描述性分析；

掌握使用热力图分析出对房屋价格影响程度最大的20个便利设施。

## 实验清单

|  |  |  |  |
| --- | --- | --- | --- |
| **实验** | **简述** | **难度** | **软件环境** |
| 短房租  分析项目 | 基于加拿大多伦多市Airbnb网站上短租房分析数据，对短租房价格变动特征、短租房价格影响因素开展数据分析任务。 | 中级 | Python 3.8.8  Jupyter Notebook |

## 实验环境介绍

### 本章实验主要使用Airbnb的短租房价格数据，对短租房的房屋信息、价格信息等等特征进行分析。主要的实验环境为Jupyter Notebook，python版本为3.8.8。

## 前置理论知识

### 可视化图形介绍

数据可视化是将数据进行图形展现，将枯燥的数据通过图形直观表示出来。此处，我们介绍本次实验所涉及到的图形的作用，便于我们加深对各种统计图形的认识与理解。

**图形1 柱状图**

柱状图（Bar Chart）：观察数据的分布情况，柱状图是一种以长方形的长度为变量的统计图表，长条图用来比较两个或以上变量的取值情况，通常利用于较小的数据集分析。柱状图亦可横向排列，或用多维方式表达。

**图形2 折线图**

折线图（Line Chart）：折线图可以显示随时间而变化的连续数据，非常适用于显示在相等时间间隔下数据的趋势。当有多个系列时，尤其适合使用折线图；当有一个系列时，应当考虑使用类别图。如果有几个均匀分布的数值标签（尤其是年），也应该使用折线图。如果拥有的数值标签多于十个，请改用散点图。

**图形3 直方图**

直方图（Histogram）：直方图又称质量分布图，是由一系列高度不等的纵向线段表示数据分布的情况。一般用横轴表示数据类型（或等成区间），纵轴表示分布情况（落在该区间的频数）。直方图的常见作用是呈现数据点波动的状态，较为直观的传递有关过程状况的信息。

**图形4 箱线图**

箱线图：箱线图属于描述性统计的一种，描述性统计的作用是用数值来描述数据有何相同与不同的地方，对数据做总结归纳。箱线图是一种使用五个数值（最小值、中位数、最大值和两个四分位数）来描述数据集分布的方法，就是观察数据集大概是集中在什么区域，分布的情况是对称的还是向左或向右偏，有没有一些数据是异常值等等。

**图形5 多变量图**

多变量图：以seaborn库中的pairplot函数举例说明，pair是成对的意思，用来展现变量两两之间的关系，包括线性/非线性关系等等。

# 短房租分析实验

## 实验总体设计

## 实验详细设计

### 材料准备

双击“短房租分析”文件夹下的“toroto”文件夹，发现有如下图2-1所示的文件夹。首先，我们要对“toroto”文件夹内的文件及其作用进行说明。

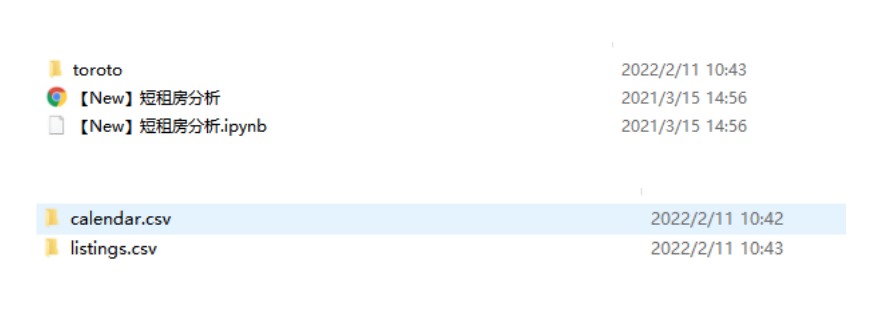


图2-1 实验文件清单

“calendar.csv”包含约650万条租房交易数据，拥有每一天每一所住房的入住信息。

“listings.csv”包含约有2万条数据，记录着所有的房屋信息、包括价格在内的几十项信息字段。

本次实验的数据分析过程都将基于这两个数据集。

### 短房租价格特征分析

#### 实验内容

本部分我们将使用“calendar.csv”对房租价格特征进行分析，首先我们导入数据表并查看数据描述。

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

%matplotlib inline

calendar = pd.read\_csv('toroto/calendar.csv.gz') #载入数据集

print('We have', calendar.date.nunique(), 'days and', calendar.listing\_id.nunique(), 'unique listings in the calendar data.')

输出如下：

We have 365 days and 17333 unique listings in the calendar data.

可以发现我们的数据时间尺度为365天，数据量为17333条。在此基础上，我们查看数据表calendar的前5行数据，使用calendar.head()函数。

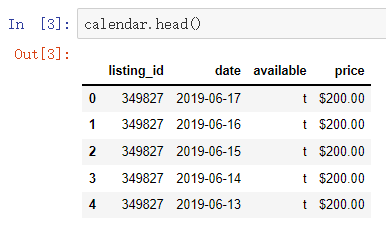


图2-2 calender数据集特征描述

接下来，我们查看数据的日期范围，以及特征的基本信息。

calendar.date.min(), calendar.date.max()

calendar.isnull().sum()

calendar.shape

calendar.available.value\_counts()

结果如下：

('2018-10-06', '2019-10-05')

listing\_id 0

date 0

available 0

price 4069436

dtype: int64

(6326545, 4)

f 4069436

t 2257109

Name: available, dtype: int64

其中，f代表已经被租用，t代表可以被出租。

接下来，我们对calendar的日期格式进行转换，并计算“busy”字段进行入住率的计算，计算方法：对available特征取值进行判断，取值为t时记为0，取值为f时记为1。

calendar\_new = calendar[['date', 'available']]

calendar\_new['busy'] = calendar\_new.available.map( lambda x: 0 if x == 't' else 1)

#计算busy特征，将t-f转换为0-1

calendar\_new = calendar\_new.groupby('date')['busy'].mean().reset\_index()

#计算新特征busy以日期分组输出，并计算busy的平均值，同时更改索引。

calendar\_new['date'] = pd.to\_datetime(calendar\_new['date'])

#将calendar\_new中的date转换为datetime格式

calendar\_new.head()

结果如下：

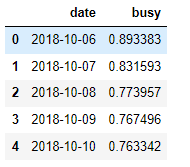


图2-3 短租房入住率结果

基于入住率busy特征，我们对房租效果进行可视化分析：

plt.figure(figsize=(10, 5)) #制定画布

plt.plot(calendar\_new['date'], calendar\_new['busy']) #绘制date时间和busy入住率数据

plt.title('Airbnb Toronto Calendar') #绘制图形的标题

plt.ylabel('% busy') #绘制ylabel

plt.show(); #绘制图形

输出如下：

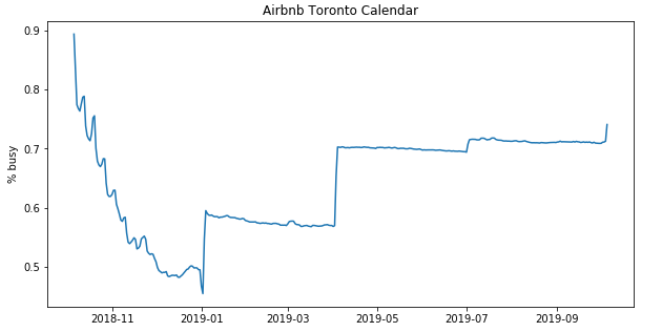


图2-4 短租房入住率可视化

通过图2-4，我们可以发现10-11月是入住率较低的阶段，2019年的7-10月较为火热，整个短租房市是在下半年较为火热。

接下来，我们查看一年当中价格的走势变化。由于原有的数据中价格部分带有$符号和,号，我们首先需要对数据进行清晰与时间字段的转换，处理好时间字段后，我们使用柱状图进行数据可视化。

calendar['date'] = pd.to\_datetime(calendar['date']) #转换日期格式

calendar['price'] = calendar['price'].str.replace(',', '') #使用空值替换,

calendar['price'] = calendar['price'].str.replace('$', '') #使用空值替换$

calendar['price'] = calendar['price'].astype(float) #转换price的格式

calendar['date'] = pd.to\_datetime(calendar['date']) #转换日期格式

mean\_of\_month = calendar.groupby(calendar['date'].dt.strftime('%B'),sort=False)['price'].mean()

mean\_of\_month.plot(kind = 'barh' , figsize = (12,7))

plt.xlabel('average monthly price'); #增加xlabel

输出如下：

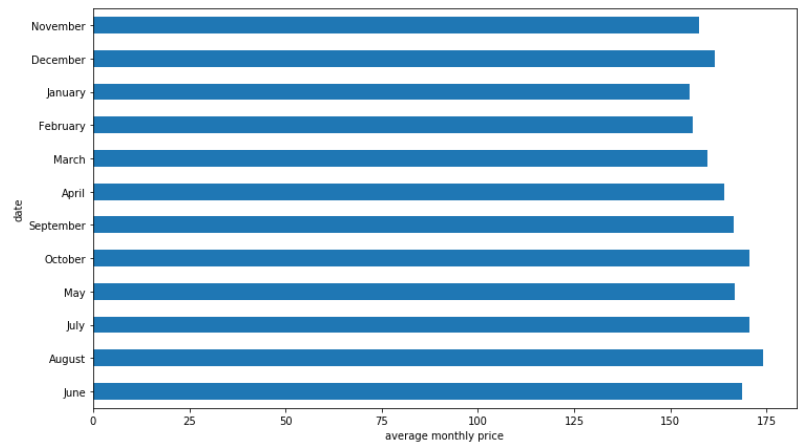


图2-5 月平均价格的柱状图可视化

可以发现，6、8、10月的是平均房价最高的三个月。接下来，我们尝试将时间跨度改为每天，看看在一周内的七天里，这些短租房的价格走势如何。

calendar['dayofweek'] = calendar.date.dt.weekday\_name

cats = [ 'Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday', 'Sunday']

price\_week = calendar[['dayofweek','price']] #提取出两列

price\_week = calendar.groupby(['dayofweek']).mean().reindex(cats) #输出dayofweek的均值列

price\_week.drop('listing\_id', axis=1, inplace=True) #剔除listing\_id列

price\_week.plot()

ticks = list(range(0, 7, 1)) #设置坐标轴的刻度

labels = "Mon Tues Weds Thurs Fri Sat Sun".split()

plt.xticks(ticks, labels);

输出如下：

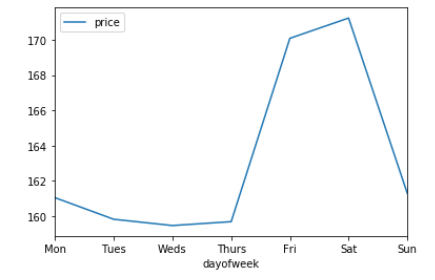


图2-6 星期x的价格波动折线图

同我们的预料一致，短租房大多是为了旅游而存在，所以每周五周六的价格都比其他时间贵一些。周末双休，使得入住的时间是周五周六的两个晚上。

#### 实验小结

本节是对短租房整体价格进行分析，从月、日尺度对价格进行描述性与可视化分析，使用到的是折线图和柱状图。通过上述分析，我们可以发现每年的6-10月是入住率较高，即租房情况较为火热的时间段，其中6、8、10是平均价格最高的三个月；从日尺度来看，星期五和星期六的价格是较高的，所得结论符合实际情况。

### 短房租价格影响因素分析

#### 实验内容

在上一节分析房屋价格波动情况的基础上，本节我们对房价的影响因素及相关性进行分析与挖掘。

数据集“listings”中记载短租房房屋的基本信息，我们查看房屋的基本信息以及房屋临近的街区，结果如图2-7，反映的是房屋临近街区的信息。

listings = pd.read\_csv('toroto/listings.csv.gz') #读取数据

print('We have', listings.id.nunique(), 'listings in the listing data.') #打印数据量

listings.groupby(by='neighbourhood\_cleansed').count()[['id']].sort\_values(by='id',ascending=False).head(10) #分街区汇总

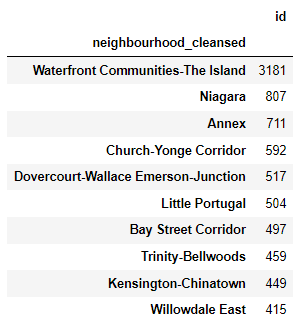


图2-7 房屋临街信息及数量

将房屋信息读取完毕后，我们对房屋的评价得分进行探索，对住房评价进行整体分析，可视化效果如图2-8所示。

plt.figure(figsize=(12,6)) #设置画布大小

sns.distplot(listings.review\_scores\_rating.dropna(), rug=True) #绘制图形

sns.despine()

plt.show();

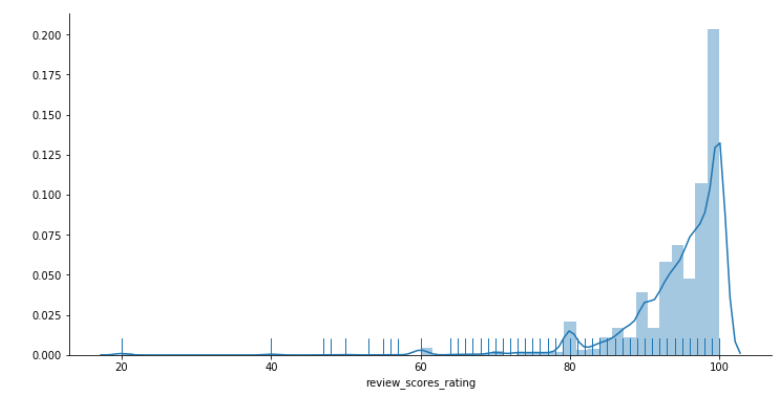


图2-8 房屋信息的整体评分情况

此处的评分是0-100分机制，可以发现房屋整体的好评率是非常高的，大多集中在80-100分之间。

接下来，我们主要分析房屋的价格状况，分析影响价格的因素及其影响程度。首先，我们对短租房的整体价格进行描述性分析，代码及结果如下：

listings['price'] = listings['price'].str.replace(',', '') #将,替换为空格

listings['price'] = listings['price'].str.replace('$', '') #将￥替换为空格

listings['price'] = listings['price'].astype(float)

listings['price'].describe() #对价格进行描述统计分析

输出如下：

count 17343.000000

mean 144.363259

std 245.723788

min 0.000000

25% 65.000000

50% 100.000000

75% 166.000000

max 12933.000000

Name: price, dtype: float64

可以发现短租房的平均价格为144.36，最小值为0，最大值为12933，房价总计为17343。此外，可以发现房价的差距非常大，最昂贵的房价为12933/晚，搜索资料发现该房屋是一个收藏家阁楼，该异常点拉高了整体的房价状况，我们通过listings.iloc[np.argmax(listings['price'])]完成定位。在数据分析任务中，我们需要服从正态分布的原则，对于这种极端情况的而存在，我们需要进行清理，因此需将价格为0和超过600的价格数据进行过滤。

listings.loc[listings['price'] > 600]['price'].describe() #查看价格超过600元的数据的描述性分析

listings.loc[listings['price'] == 0]['price'].count() #查看价格为0元的数据

plt.figure(figsize=(12,6))

listings.loc[(listings.price <= 600) & (listings.price > 0)].price.hist(bins=200) #对筛选后的数据进行直方图可视化

plt.ylabel('Count')

plt.xlabel('Listing price in $')

plt.title('Histogram of listing prices');

输出如下：

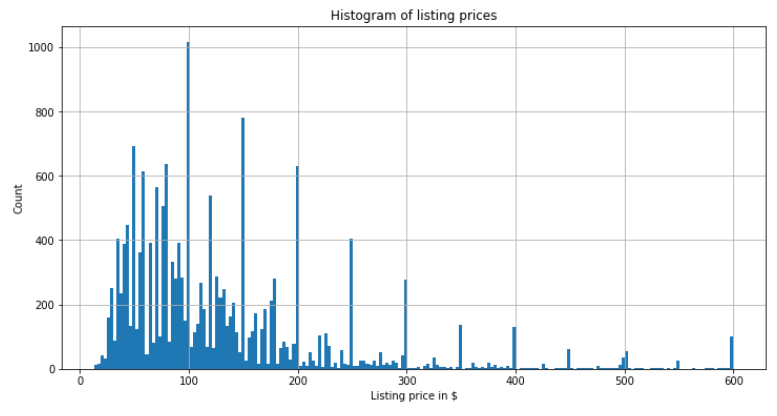


图2-9 房屋价格分布直方图

通过图2-9的直方图，可以发现房价大多集中在100附近，房价超过200的房屋数量较少。

接下来，我们观察Superhost这个字段，反应的是某房屋是否为高级房。高级房是指满足一定的评级要求，比如100次以上成功的预订交易，好评率超过90%等等。我们分析这个评级的房屋与不具备该评级的房屋在价格上是否存在差别。

sns.boxplot(y='price', x='host\_is\_superhost', data=listings.loc[(listings.price <= 600) & (listings.price > 0)]) #绘制箱线图，数据筛选的是房价在0到600之间的信息 plt.show();

输出如下：

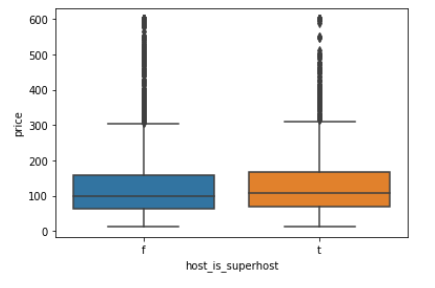


图2-10 高级房房价与普通房对比

可以发现，高级房的价格会略高于普通房。

接下来，我们分析房屋的装修特性和价格的关系，通过箱线图判断不同的装修特性和价格的关系。

plt.figure(figsize=(18,10)) sort\_price = listings.loc[(listings.price <= 600) & (listings.price >0)]\.groupby('property\_type')['price'].median().sort\_values(ascending=False).index sns.boxplot(y='price', x='property\_type', data=listings.loc[(listings.price <= 600) & (listings.price > 0)], order=sort\_price) ax = plt.gca() ax.set\_xticklabels(ax.get\_xticklabels(), rotation=45, ha='right') plt.show();

输出如下：

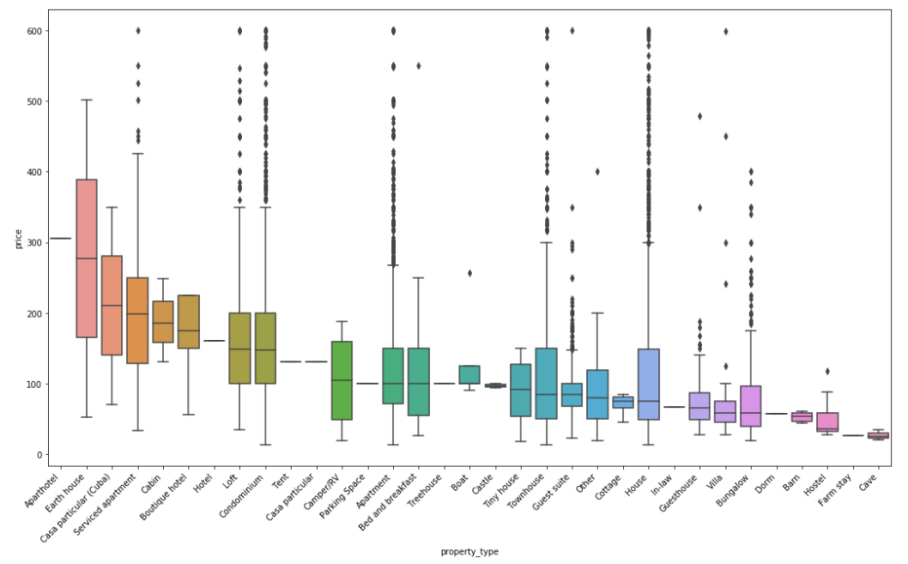


图2-11 不同装修风格对价格的影响对比

类似的，我们接着探究房型对价格的影响：

sort\_price=listings.loc[(listings.price<=600)&(listings.price >0)]. groupby('room\_type')['price'].median(). sort\_values(ascending=False).index sns.boxplot(y='price', x='room\_type', data=listings.loc[(listings.price <= 600) & (listings.price > 0)], order=sort\_price) ax = plt.gca() ax.set\_xticklabels(ax.get\_xticklabels(), rotation=45, ha='right') plt.show(); listings.loc[(listings.price <= 600) & (listings.price > 0)].pivot(columns = 'room\_type', values = 'price').plot.hist(stacked = True, bins=100) plt.xlabel('Listing price in $');

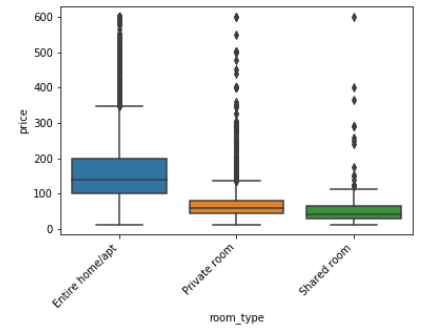


图2-12 房型-房价的箱线图

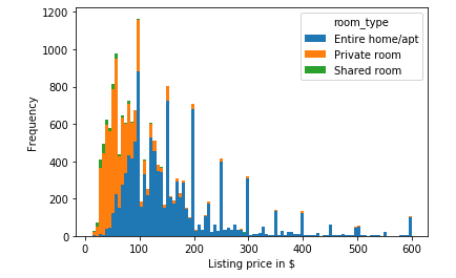


图2-13 房型-房价的堆积条形图

可以发现，租整套的价格明显比多人合租要贵一些。此外，观察图2-13可以发现，在房价为100元一下的房屋中，单间的房屋较多；在100-200元及200元以上的房屋中，整套房的房屋较多。

此外，由于每间房内的设施状况都不同，我们分析下影响房价中较为重要的便利设施，并探索这些便利设施与价格间的关系。

listings['amenities'].head() listings.amenities = listings.amenities.str.replace("[{}]", "").str.replace('"', "") #替换字符 listings['amenities'].head() #查看数据头部

输出如下：

0 TV,Internet,Wifi,Air conditioning,Kitchen,Indo...

1 TV,Cable TV,Internet,Wifi,Air conditioning,Poo...

2 Wifi,Air conditioning,Kitchen,Free parking on ...

3 Internet,Wifi,Pets live on this property,Cat(s...

4 Internet,Wifi,Air conditioning,Kitchen,Free pa...

Name: amenities, dtype: object

pd.Series(np.concatenate(listings['amenities'].map(lambda amn:amns.split(",")))).

value\_counts().head(20).plot(kind='bar')

ax = plt.gca()

ax.set\_xticklabels(ax.get\_xticklabels(), rotation=45, ha='right', fontsize=12)

plt.show();

输出如下：

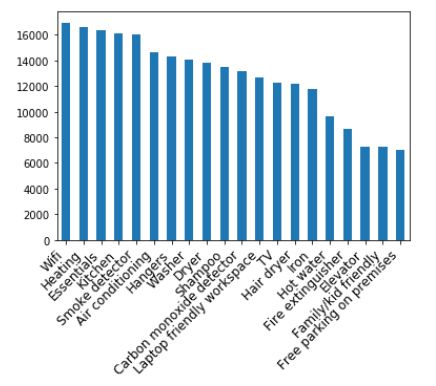


图2-14 房屋重要设施探索

可以看出，Wifi、暖气、厨房等便利设施是最重要的部分。接下来，我们来看一下前20个便利设施和价格间的关系。

amenities = np.unique(np.concatenate(listings['amenities'].map(lambda amns: amns.split(","))))

amenity\_prices = [(amn, listings[listings['amenities'].map(lambda amns: amn in amns)]

['price'].mean()) for amn in amenities if amn != ""]

amenity\_srs = pd.Series(data=[a[1] for a in amenity\_prices], index=[a[0]

for a in amenity\_prices])amenity\_srs.sort\_values(ascending=False)[:20].plot(kind='bar')

ax = plt.gca()

ax.set\_xticklabels(ax.get\_xticklabels(), rotation=45, ha='right', fontsize=12)

plt.show();

输出如下：

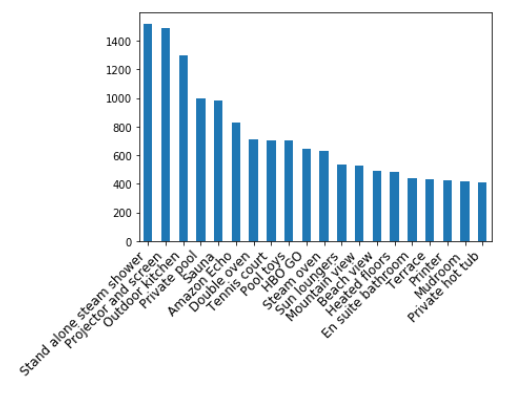


图2-15 影响房屋价格的设施重要性排序

每间房屋中床的数量也会影响房价，此处我们着重探讨床的数量对房价的影响。首先，我们从listings.columns中判断出与床相关的特征，包括host\_listings\_count、accommodates等，并从整体数据表中将其筛选出绘制Pairplot，查看多种特征之间的相关性。

col = ['host\_listings\_count', 'accommodates', 'bathrooms', 'bedrooms''beds','price', 'number\_of\_reviews','review\_scores\_rating', 'reviews\_per\_month']

sns.set(style="ticks", color\_codes=True)

sns.pairplot(listings.loc[(listings.price <= 600) & (listings.price > 0)][col].dropna())

plt.show();

输出如下：

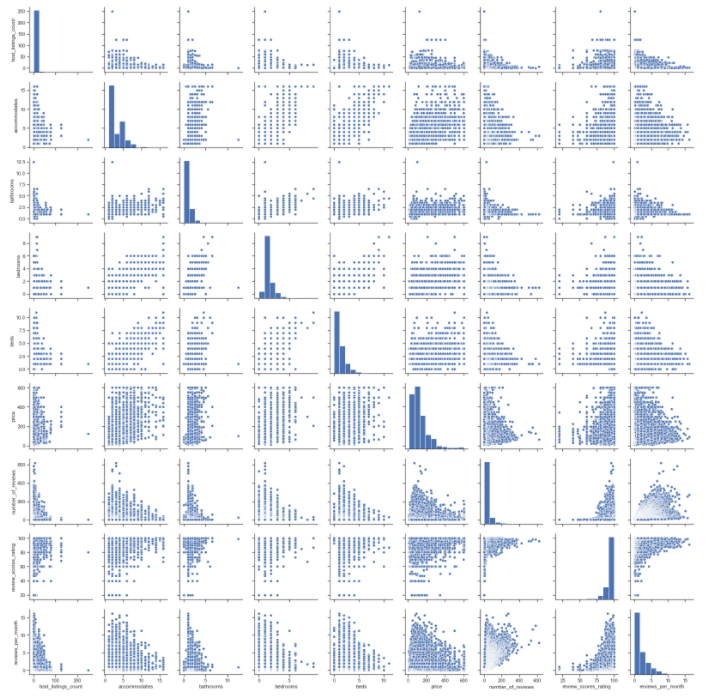


图2-16 房屋与床相关的特征间的矩阵散点图

最后，我们分析价格在0-600之间的房屋各项特征之间的相关性，并绘制热力图进行可视化分析。

plt.figure(figsize=(18,10))

corr = listings.loc[(listings.price <= 600) & (listings.price > 0)][col].dropna().corr() #筛选数据，并计算相关系数矩阵

plt.figure(figsize = (6,6)) #设置画布大小

sns.set(font\_scale=1)

sns.heatmap(corr, cbar = True, annot=True, square = True, fmt = '.2f', xticklabels=col, yticklabels=col) #绘制热力图

plt.show();

输出如下：

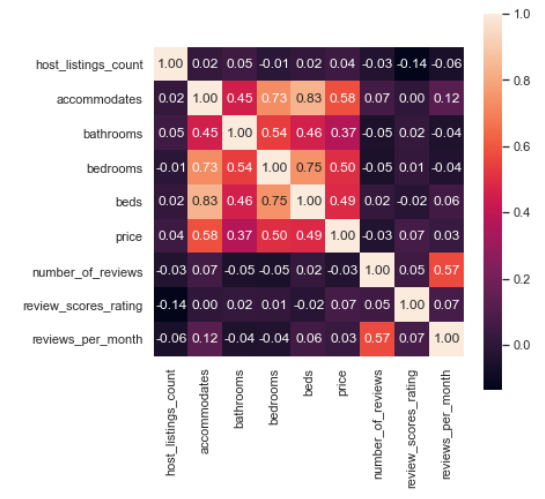


图2-17 房屋床相关信息的特征热力图

plt.figure(figsize=(18,10))

sns.heatmap(listings.loc[(listings.price <= 600) & (listings.price > 0)].groupby(['bathrooms', 'bedrooms']).count()['price']

.reset\_index().pivot('bathrooms', 'bedrooms', 'price').sort\_index(ascending=False),

cmap="Oranges", fmt='.0f', annot=True, linewidths=0.5) #设置图像相关参数

plt.show();

输出如下：

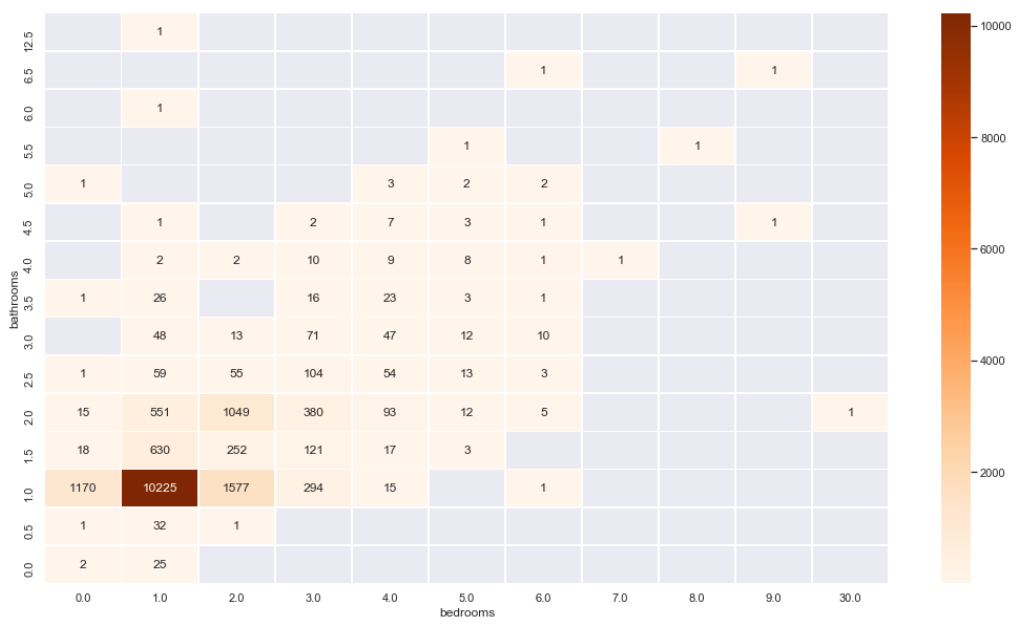


图2-18 房屋床数与洗漱间数量对应价格可视化

#### 实验小结

##### 本节实验内容中，我们在分析房屋评价状况的基础上，着重分析了影响房屋价格的特征及其影响程度。首先，从分析房价整体波动状况入手，通过绘制直方图分析房屋价格，发现大部分房屋价格分布在50-200元之间。为了得到更为准确的房屋价格信息，我们通过数据筛选与数据清洗，筛选出价格位于0-600元之间的房屋及其信息供后续分析。

我们着重分析了高级房属性对房价的影响、房屋的装修类型对房价的影响、房型对房价的影响、便利设施对房价的影响、床的相关特征对房价的影响等等。具体的分析结果请参考实验内容。

# 实验总结

## 实验总结

本实验基于多种数据可视化图形对短租房房价影响因素进行分析，主要是用到的数据集包括房价信息、房屋内部属性等多种特征。通过使用折线图分析房屋价格的走势，并按照月和日的不同尺度将波动规律展现出来：租房的客户量在每年7-9月之间较多，每年的6、8、10月，每周的周五周六两天房价较高。有关房屋信息的数据分析中，可以发现房屋的好评率较高，房屋评分大多在80分以上，房屋价格集中在0-200元的区间中。根据房屋类别的不同，可以发现高级房比普通房的房屋价格略高，整租房的价格比合租房要高。此外，通过条形图等可视化方法，我们找出前20个最重要的便利设施，拥有这些便利设施的房价通常租房价格较高；房屋内与床相关的特征也会影响价格，绝大部分房屋中包含一张床，通过矩阵散点图完成了房屋内部多种特征之间的相关性分析，最后基于此进行热力图的可视化分析，卓着重探究了房屋内床的数量、洗漱间的数量与价格之间的关系。

通过此实验，我们能够了解如何分析房屋价格特征的变化情况，找到影响房屋价格的主要因素，并对这些特征间的变化情况、构成情况、趋势情况等等进行分析。通过此学习过程，加深了我们对可视化方法的理解与运用。