姓名	学号	Github 账号	分工	成绩
羊艺超	201592041	yyc123xn	组长	
刘威	201592102	Damojinshui	组员	
欧阳诚	201592208	dlutoyc	组员	
程义序	201592350	wincyx	组员	
谢因坦	201592468	sherlockhumorous	组员	

开源软件基础课程报告

报 告 题 目　Airbnb 基于 python 的数据分析报告

项 目 网 址　https://github.com/yyc123xn/Python.github.io

完 成 日 期　2018.1.8

大连理工大学软件学院

目　　录

1 系统分析

1.1 需求分析

Airbnb 作为一个空闲房屋短期租赁平台,以一种个性化和更具特色的外出居住方式让旅行变得饱满丰富，也是共享经济接下来可以优惠的突破口，而 Airbnb 自主开发了一套叫“Aerosolve”的机器学习平台。这个平台会自动将城市划分成无数个由微型街区组成的小区域，并分析房主们拍摄的房间照片。Aerosolve 还模仿酒店和航空公司的定价模式搭建了一套动态定价策略。所以为将不同城市的房屋价格进行比较，来提供房屋价格与城市的关系，以及不同城市价格之间的比较，我们团队选定了 Airbnb 上海、北京及成都的房屋信息来进行数据分析，将不同城市的房价用图像化来进行展示和分析，并且对不同城市的数据进行了比较。

1.2 技术简介

团队使用 requests 库来爬取 Airbnb 网页数据,requests 支持 HTTP 连接保持和连接池，支持使用 cookie 保持会话，支持文件上传，支持自动确定响应内容的编码，支持国际化的 URL 和 POST 数据自动编码。

选择 MySQL 建立数据库，pymsql 是在 Python3.x 版本中用于连接 MySQL 服务器的一个库，pymsql 遵循 Python 数据库 API v2.0 规范，并包含了 pure-Python MySQL 客户端库。来存储爬到的三个城市 Airbnb 房屋信息。

选择了 selenium 的 webdirver 模块来作为网页驱动，在爬取数据的时候以动态化的方式来监控爬取过程和出现的错误。

通过 pymysql 数据接口来在 python 中操作数据库。

使用 plotly 来在线制作图表。

2 概要设计

2.1 数据库设计

2.1.1 airbnb_beijing

以 id 为主键，同时还有 title、url、price、type、comments 等属性。具体属性设定如表所示。

列名	数据类型	允许为空	说明
id	Int	不允许	房屋号，主键，自增
Title	varchar	不允许	房屋名称
url	Vervhar	不允许	网址
Price	Int	不允许	价格
Type	varchat	不允许	房屋类型
Comments	varchar	不允许	使用年限

2.2 函数设计

采用 python 库函数

2.2.1 爬取网站数据的实现

```
for i in range(0, 17 , 1):
url=
('https://zh.airbnb.com/s/'+cityUrl+'&allow_override%5B%5D=&section_offset=
{}').format(str(i))
    if i == 0:
        url = 'https://zh.airbnb.com/s/'+cityUrl+'&allow_override%5B%5D='
    driver.get(url)


    time.sleep(2)
    html = driver.page_source#获取网页的 html 数据、
    soup = bs4.BeautifulSoup(html,'lxml')#对 html 进行解析
    r = soup.find_all('div', class_='_v72lrv')
    sql = "INSERT INTO Airbnb_" + city + "(url,price,type,comments,score)
values('" + url + "','" + price + "','" + type + "','" + comments + "','" + score
+ "')"
    cursor.execute(sql)
    con.commit()
```

2.2.2 图表展示实现

```
import plotly
plotly.tools.set_credentials_file(username='yyc',
api_key='1puNznRHHWs5kPei8ESP')
import plotly.plotly as py
from plotly.graph_objs import *
```

2.2.3 城市进行比较实现

```
sql=
"SELECT Airbnb_Shanghai.type,AVG(Airbnb_Shanghai.price),AVG(Airbnb_Chengdu.
price),AVG(Airbnb_Beijing.price) FROM Airbnb_Shanghai "
        "JOIN Airbnb_Chengdu ON Airbnb_Shanghai.type = Airbnb_Chengdu.type " \
        "JOIN Airbnb_Beijing ON Airbnb_Shanghai.type = Airbnb_Beijing.type " \
        "GROUP BY type"
```

2.2.4 城市分数比较实现

```
sql = " SELECT (SELECT COUNT(*) FROM airbnb_"+city+" AS "+city+" WHERE
"+city+".score = 3) AS '3' ," \
          "IFNULL((SELECT AVG(price) FROM airbnb_"+city+" AS "+city+" WHERE
"+city+".score = 3),0) AS '3Price' ," \
          "(SELECT COUNT(*) FROM airbnb_"+city+" AS "+city+" WHERE
"+city+".score = 4) AS '4' ," \
          "IFNULL((SELECT AVG(price) FROM airbnb_"+city+" AS "+city+" WHERE
"+city+".score = 4),0) AS '4Price' ," \
          "(SELECT COUNT(*) FROM airbnb_"+city+" AS "+city+" WHERE
"+city+".score = 4.5) AS '4.5' ," \
          "IFNULL((SELECT AVG(price) FROM airbnb_"+city+" AS "+city+" WHERE
"+city+".score = 4.5),0) AS '4.5Price' ," \
```

```
            "(SELECT  COUNT(*)  FROM  airbnb_"+city+"  AS  "+city+"  WHERE
"+city+".score = 5) AS '5' ," \
            "IFNULL((SELECT AVG(price) FROM airbnb_"+city+" AS "+city+" WHERE
"+city+".score = 5),0) AS '5Price' " \
              "FROM airbnb_"+city+" LIMIT 1 "
```

2.2.5 计算各个评分段的堆叠占比实现

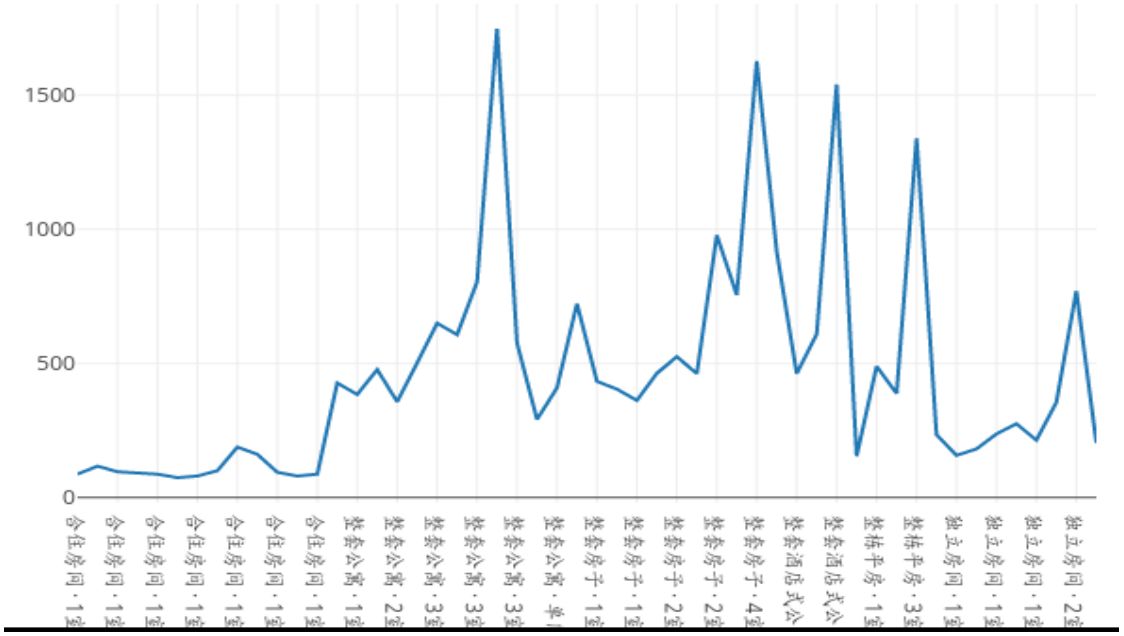
```
sql = " SELECT (SELECT COUNT(*) FROM airbnb_"+city+" AS "+city+" WHERE
"+city+".score = 3) AS '3' ," \
            "IFNULL((SELECT AVG(price) FROM airbnb_"+city+" AS "+city+" WHERE
"+city+".score = 3),0) AS '3Price' ," \
            "(SELECT  COUNT(*)  FROM  airbnb_"+city+"  AS  "+city+"  WHERE
"+city+".score = 4) AS '4' ," \
            "IFNULL((SELECT AVG(price) FROM airbnb_"+city+" AS "+city+" WHERE
"+city+".score = 4),0) AS '4Price' ," \
            "(SELECT  COUNT(*)  FROM  airbnb_"+city+"  AS  "+city+"  WHERE
"+city+".score = 4.5) AS '4.5' ," \
            "IFNULL((SELECT AVG(price) FROM airbnb_"+city+" AS "+city+" WHERE
"+city+".score = 4.5),0) AS '4.5Price' ," \
            "(SELECT  COUNT(*)  FROM  airbnb_"+city+"  AS  "+city+"  WHERE
"+city+".score = 5) AS '5' ," \
            "IFNULL((SELECT AVG(price) FROM airbnb_"+city+" AS "+city+" WHERE
"+city+".score = 5),0) AS '5Price' " \
              "FROM airbnb_"+city+" LIMIT 1 "
```

3 分析及分析结果

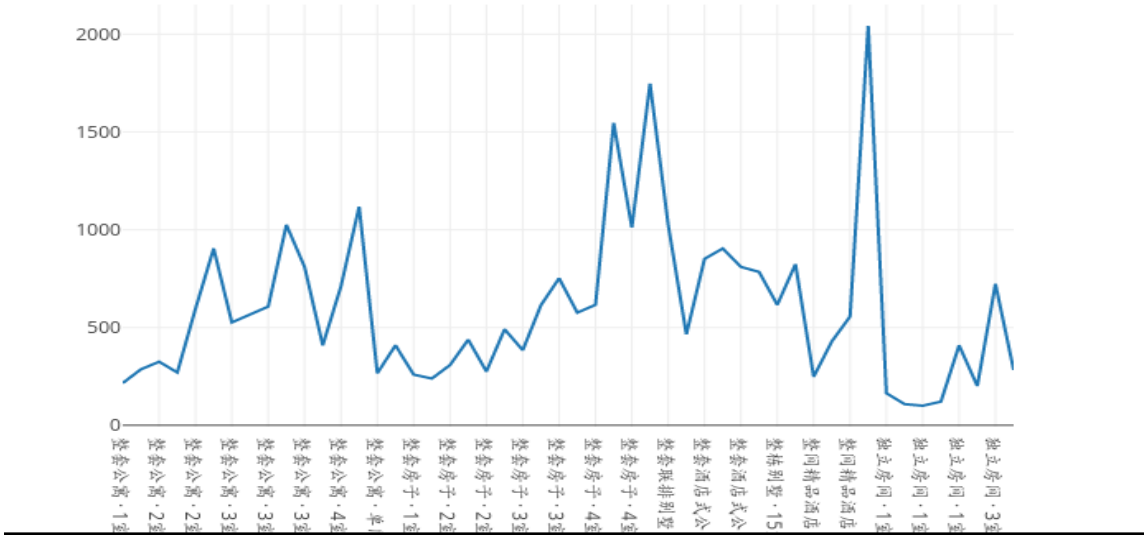
3.1 房价及型号分析图

三座城市房价及房屋型号分析图如下

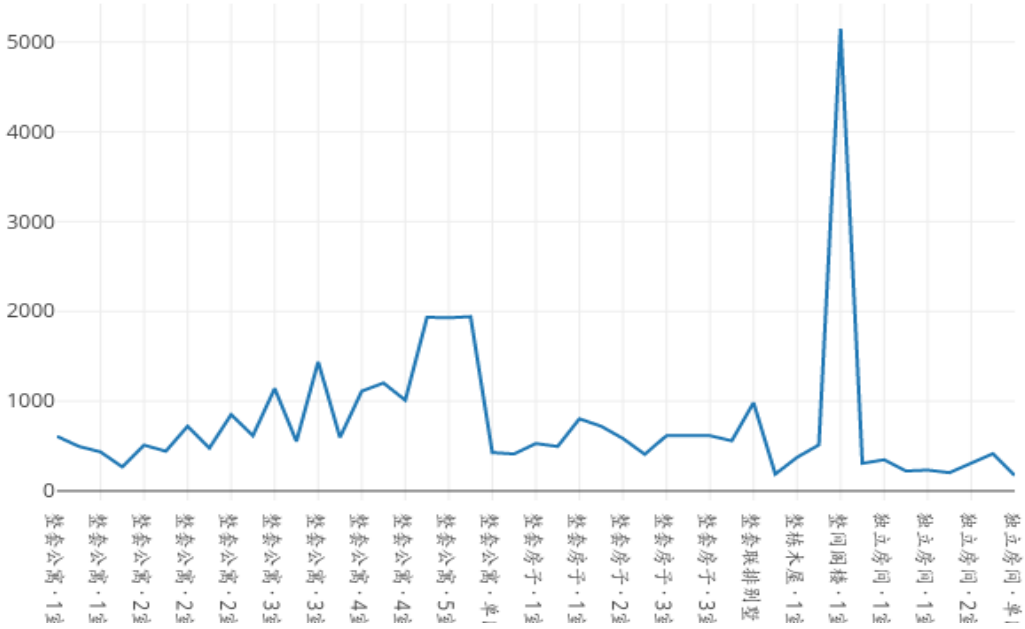
北京：



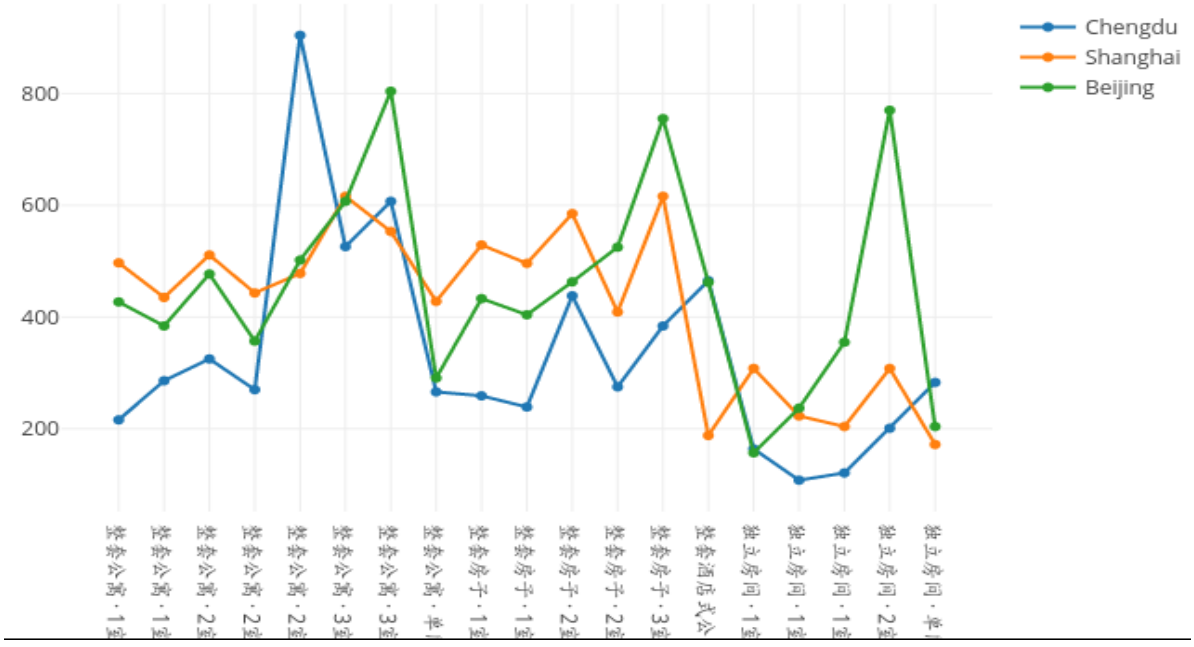
成都：



上海：

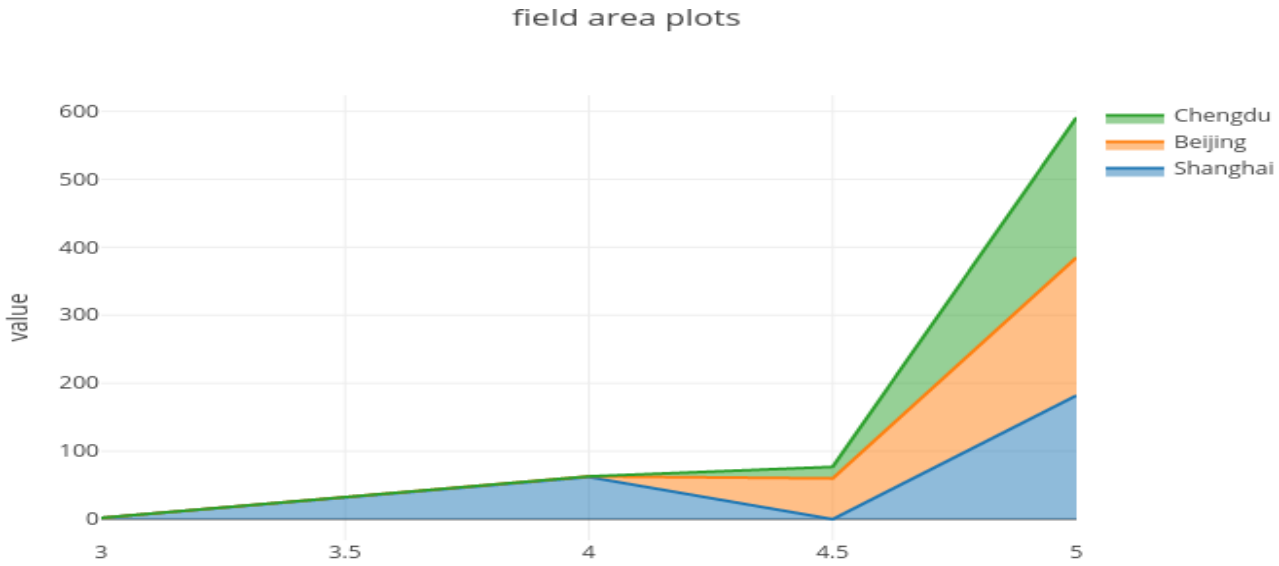


三座城市综合比较：

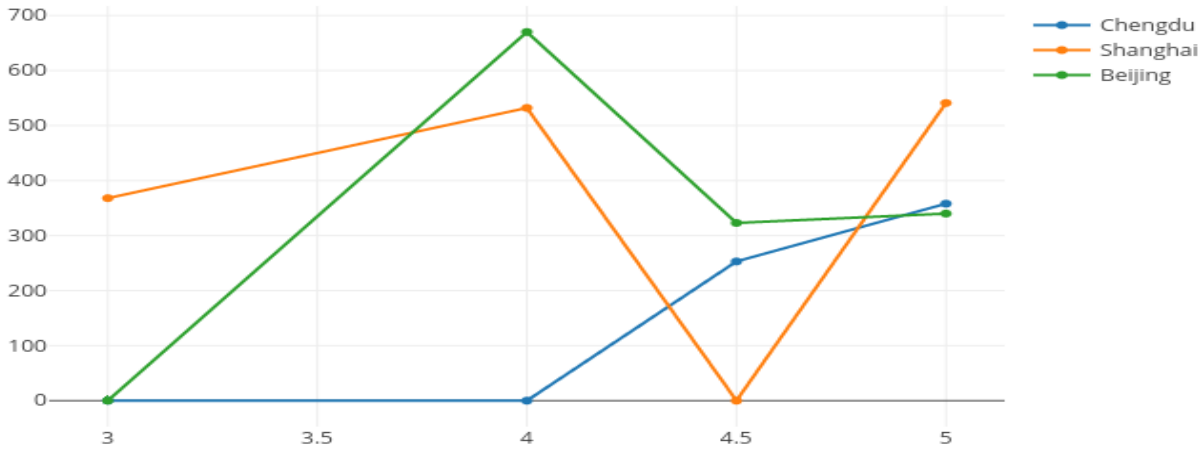


3.2 价格区域评分图

我们将三个城市的房屋的评分与平均房价进行纵向对比，通过计算不同价格房屋在不同评分段的堆占比分析房价与房屋质量之间的关系

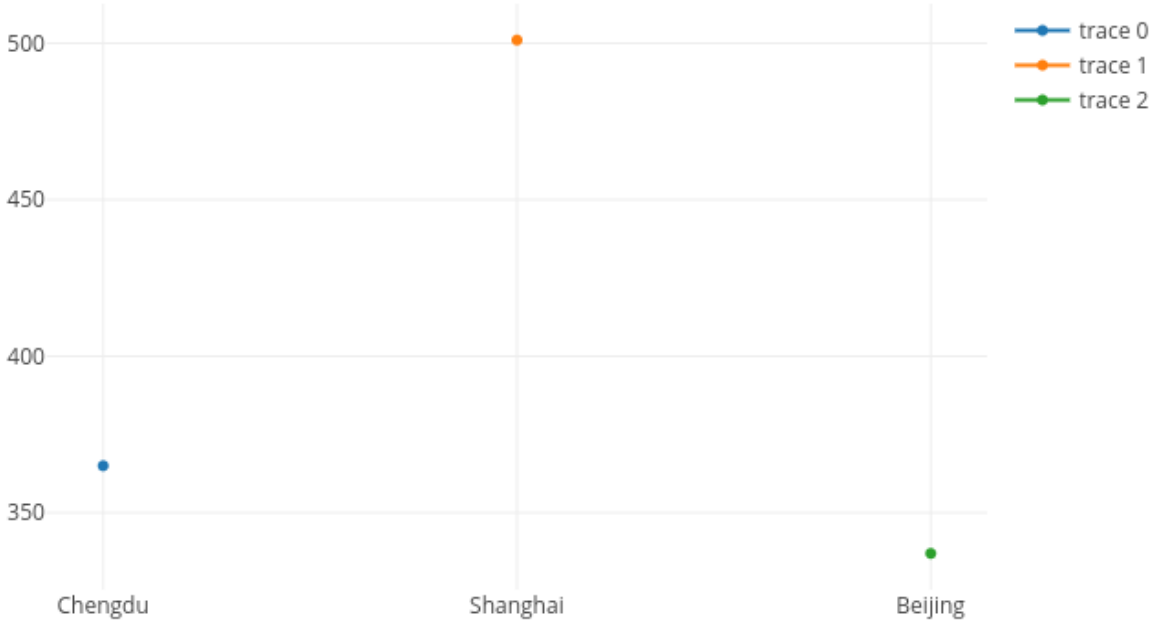


结论：上图说明在 5 分的评价中成都房屋占比最高，而在 4 分评价中，三者的房屋价格基本一致，说明 Airbnb 中房屋还是和质量相对成正比的。



结论：由上海成都北京三座城市的曲线，从而分析得出上海的 Airbnb 的房价在同等质量下价格偏高，而北京的房屋在 4 分评价中价格很高，而成都的房价评分普遍偏高，并且价格也相对较低。

3.3 房价均值图



结论：从三座城市的房价均值中，可以看出上海的房价均值最高。

4 总结

这次的作业有很多不足之处，但是我们在这次大作业中应用了很多开源基础上的知识，使我们对开源有了更深层次的了解，希望接下来有机会继续学习开源知识。

本次大作业在我们进行分析之后应该进一步给出完善措施，通过不同城市的大数据分析得到的结论来动态调节 Airbnb 不同城市短租型房屋的价格，后期我们会在此基础上继续学习开发动态定价策略，争取完善出自己小组的动态定价策略。