Boston-Airbnb-data analysis

目录

一. 数据集简介.....2
二. 研究问题.....2
三. 结果分析.....2
Analyse-1 2
Analyse-2 4
Analyse-3 4
Analyse-4 6
Analyse-5 7

一. 数据及简介

波士顿 Airbnb 公开数据是共享民宿网站 Airbnb 的开放数据，包括在波士顿地区的民宿列表(listings.csv)、不同时间的价格(calendar.csv)、用户评分及评论(reviews.csv)。

- Listings.csv 的内容是对民宿的细节描述，主要包括 listings, price, accomodates, ratings, number of reviews, summary, name, owner name, Description, host Id 等。
- Calendar.csv 是对 listings 的描述，是否可预订(availability),价格(price)
- Reviews.csv 是对 Boston 的每个民宿(listing)的评价(review)。

二. 解决问题

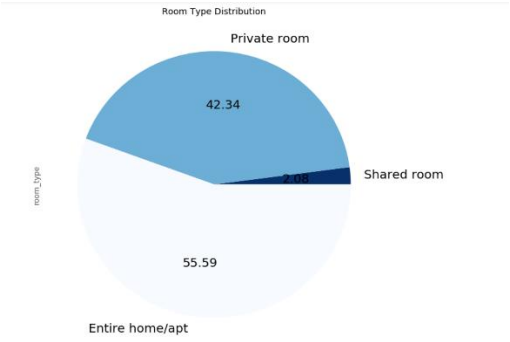
本次实验主要分析以下五个问题：

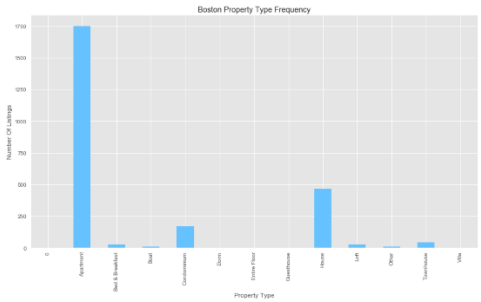
- 哪些因素造成了民宿(listings)的价格(price)的差异？
- 在 Boston 投资哪一块地产会从 Airbnb 公司得到最大的回报？
- 民宿价格的季节性变动特征。
- 分析评论和民宿价格之间的关系
- 对民宿价格进行分析和预测

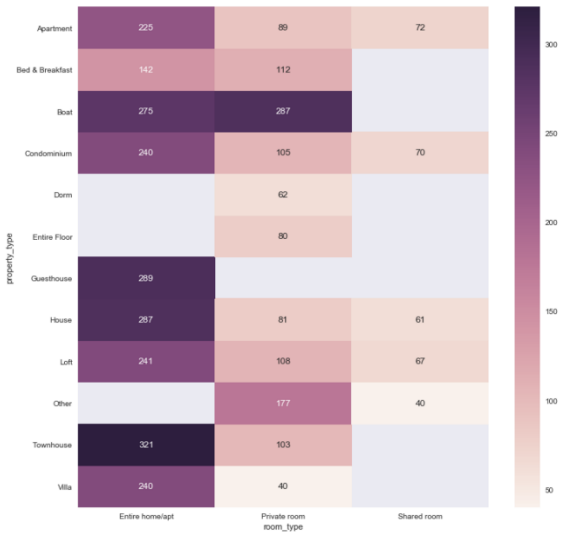
三. 结果分析

Analyse-1. 哪些因素造成了民宿(listings)的价格(price)的差异？

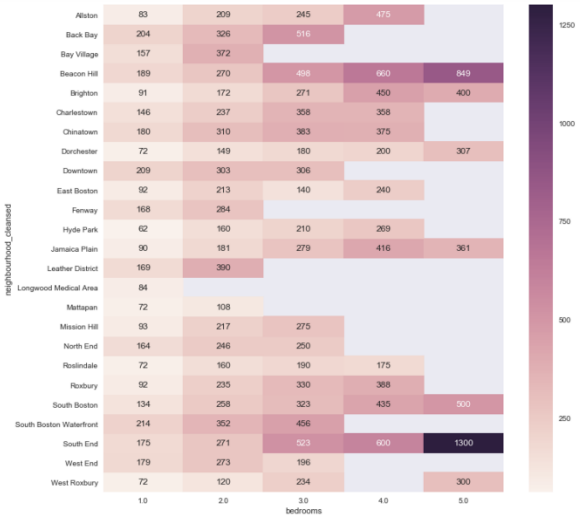
1. 对民宿列表(listing.csv)进行数据清洗后，展现了根据不同的 room_type 和 property_type 的民宿数量绘制的图表，可以得出结论，相比 private rooms 或者 shared rooms，人们更愿意选择 entire home/apt，property_type 也是一个重要因素，其中，apartment 和 home 的民宿占了绝大多数。



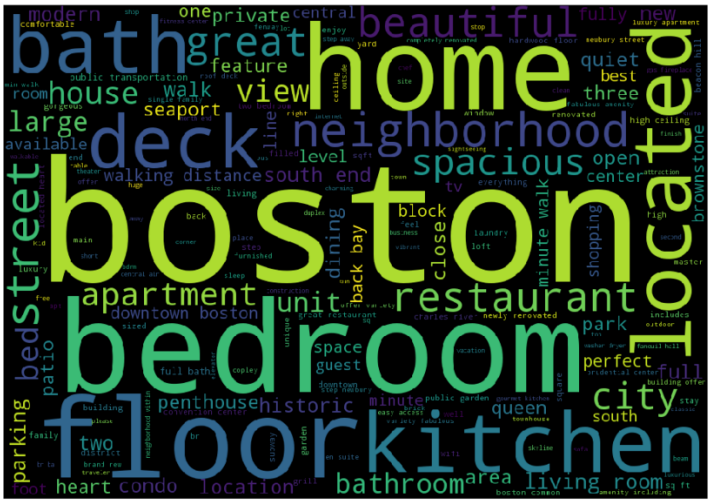




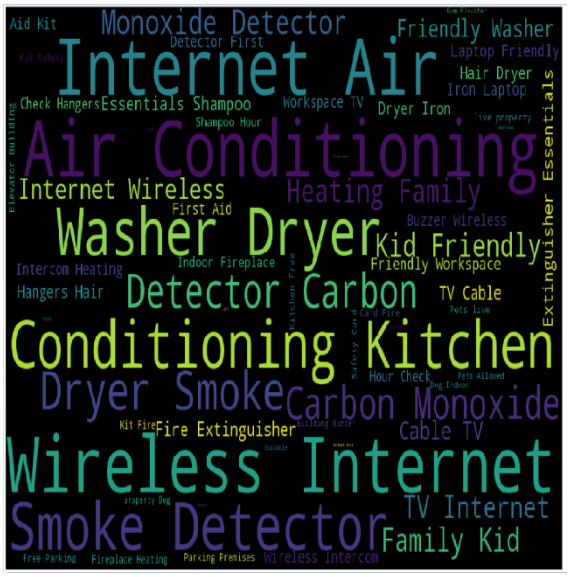
2. 绘制了不同的neighbourhood_cleansed和bedrooms下的民宿数量(listings). 从热力图中可以看出随着 bedrooms 数量的增加，民宿数量(listings)也随之增加，另外，neighbourhood_cleansed 对民宿数量(listings)也有影响。



3. 分析 summary 变量对价格(price)的影响，将价格(price)排名前 100 的民宿(listings)的 summary 用可视化结果展现出来，可以看出，home, south Boston, bedroom, floor, kitchen, restaurant, spacious, neighborhood ,located 这些关键词在描述民宿(listings)时被使用的最频繁。这一结果可能是由于 Airbnb 的主要目的不是提供豪华酒店，而是提供便利的住宿场所。



4. 分析民宿(listings)具备哪一项基础设施(amenities)更可能有更高的价格。将价格排名前 30 的民宿(listings)的基础设施(amenities)使用词云的方法展现出来，结果显示，具有 Air conditioning, washer/dryer, Kid friendly, Heating, hair dryer, buzzer 等基础设施(amenities)的民宿(listings)更可能有更高的价格。

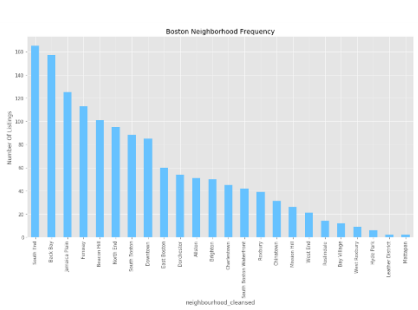


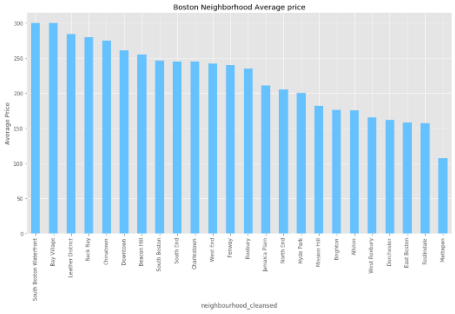
Analyse-2. 在 Boston 投资哪一块地产会从 Airbnb 公司得到最大的回报？

1. 对 listings.csv 进行数据清洗，计算不同的 room_type 的平均价格(average_Price)，可知 Entir home/apt 的平均价格最高。

	room_type	average_Price
0	Entire home/apt	232.322326
1	Private room	89.505184
2	Shared room	69.903846

2. 计算不同的社区(neighbourhood_cleansed)的民宿数量(Number_Of_Listings)和平均价格(Average_price),并使用柱状图的形式展现出来。从结果可以看出，South End，Back Bay 和 Jamaica Plain 地区的民宿数量排名较靠前；South Boston Waterfront, Bay Village 和 Back Bay 地区的民宿平均价格较高。综合来看，'Back Bay'和'South Boston'可以被认为民宿数量最多且价格较为靠前的地区。

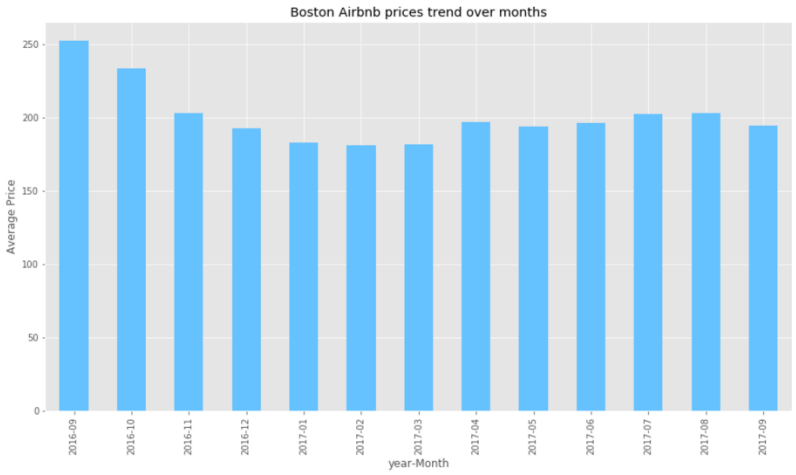




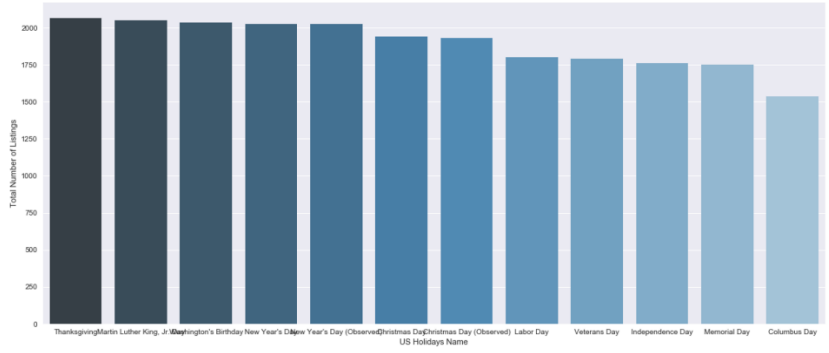
Analyse-3. 民宿价格的季节性变动特征。

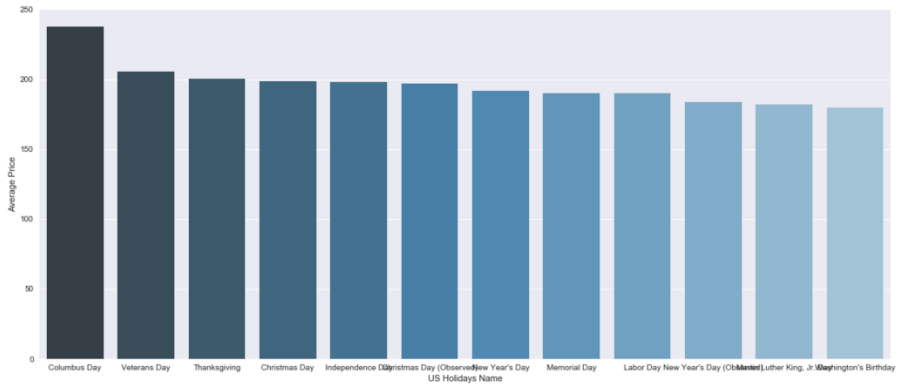
1. 对 calendar.csv 进行数据清洗，得到以下结果计算出不同月份的民宿平均价格，并且使用柱状图展现出来，从结果可以看到，2016 年 9 月和 2016 年 10 月的民宿平均价格是最高的。

	listing_id	date	available	price	Year	Month	Day
365	3075044	2017-08-22	t	65.0	2017	08	22
366	3075044	2017-08-21	t	65.0	2017	08	21
367	3075044	2017-08-20	t	65.0	2017	08	20
368	3075044	2017-08-19	t	75.0	2017	08	19
369	3075044	2017-08-18	t	75.0	2017	08	18



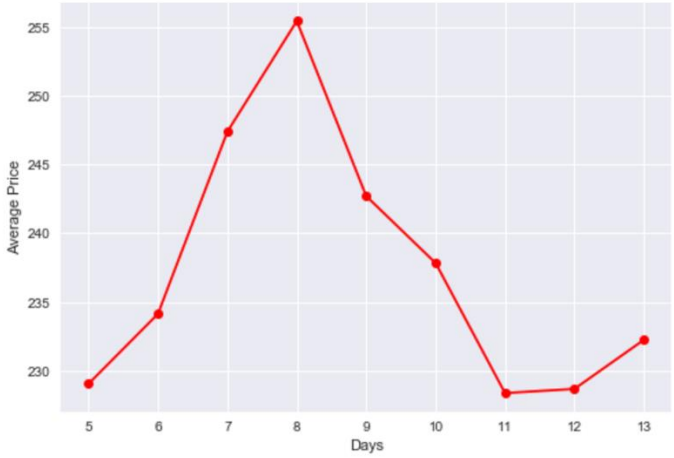
2. 分析节假日(holiday)对民宿的数量和平均价格是否有影响，并将结果用柱状图展现出来。从结果可以看出，感恩节的民宿数量是最多的，Columbus Day 的民宿价格是最高的。





3. 分析 2016 年 9 月和 2016 年 10 月这两个月份在每个工作日的价格变动特

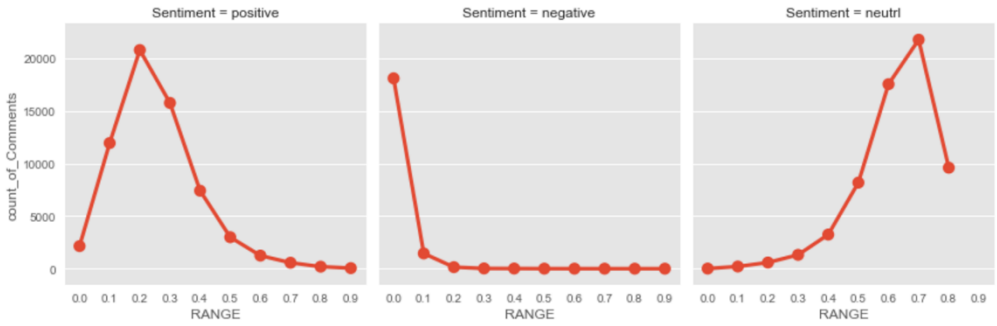
点，从结果可以看出，周末民宿的价格比工作日的价格高。



Analyse-4. 分析评论和民宿价格之间的关系

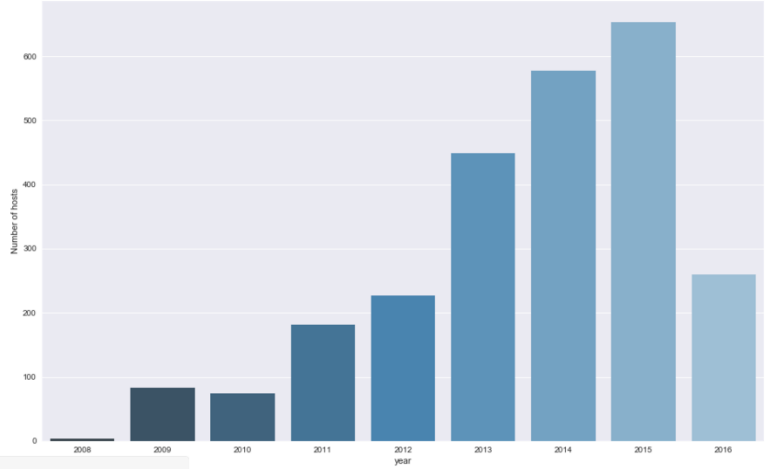
1. 对 reviews.csv 进行数据清洗，对每个评论(comments)进行情感分析得到对应 positive, negative 和 neutrl 的值，结果如下所示，再按照不同的情绪类别展现评论数量和情绪得分之间的关系。从结果可以看出，几乎没有一条评论是完全消极的，大部分的评论的消极情绪为 0，大部分的评论是中立情绪的，积极性的评论也占了很大的比例。

	listing_id	id	date	reviewer_id	reviewer_name	comments	polarity_value	neg	pos	neu	compound	language
0	1178162	4724140	2013-05-21	4298113	Olivier	My stay at islam's place was really cool! Good...	{'neg': 0.0, 'neu': 0.648, 'pos': 0.352, 'comp...	0.0	0.352	0.648	0.9626	en
1	1178162	4869189	2013-05-29	6452964	Charlotte	Great location for both airport and city - gre...	{'neg': 0.0, 'neu': 0.639, 'pos': 0.361, 'comp...	0.0	0.361	0.639	0.9061	en



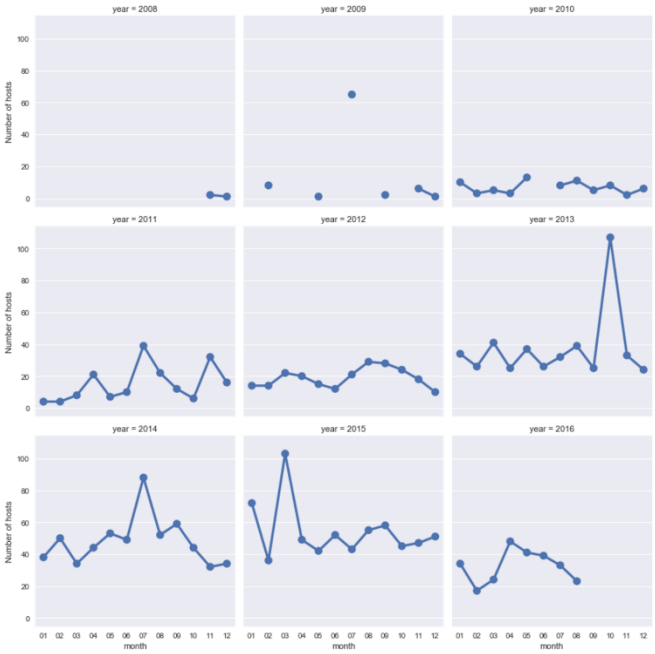
2. 对 Listings.csv 进行数据清洗，分析评论的数量和民宿的价格之间的关系，并绘制的散点图。从结果可以看出，100-400 价位的民宿获得的评论数是最多的。

3. 分析了评论中最主要涉及的内容，并用词云的形式展现出来。从结果来看，大部分的评论主要集中在"great location", "great host", "walking distance" 和 "highly recommended"这些关键词上。

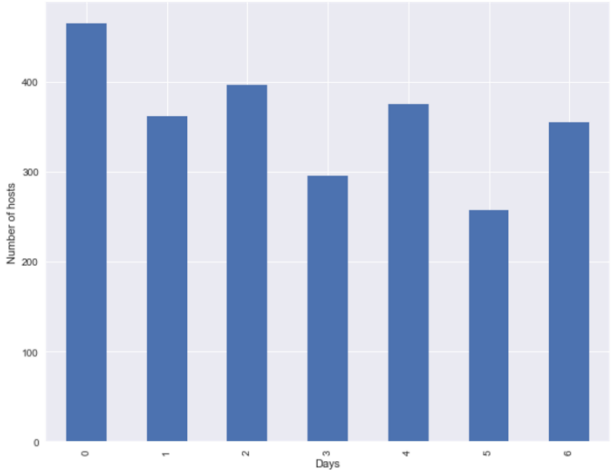


Analyse-5. 对民宿价格进行分析和预测

1. 首先用柱状图展示了每天和每年注册的民宿数量，再对每年每个月注册的民宿数量绘制折线图。结果显示，2015 年注册的民宿数量最多，且大多数房东选择在 5 月，7 月和 11 月注册。



2. 此外，分析了房东们是否更愿意在周末注册。结果显示，房东们更愿意在周末注册。



3. 建立模型对价格进行预测，使用多个回归模型对价格进行预测，从结果来看，线性回归更适合来预测该价格变动趋势。

回归模型	预测误差	模型得分
线性回归	34.35077973783336	0.5767463453171999
岭回归	34.34221939660347	0.5768952101108253
Lasso 回归	36.251902103484554	0.56516864935726
多项式回归	34.07044257583766	0.566662495785181
ElasticNet 回归	42.01092972722799	0.51024633018741

决策树回归	34.786885245901644	0.5444245742799942
逻辑回归	94.0	0.03589232303090728