2022

نویسندگان	عنوان
علیرضا افروزی، محمد کربلایی شعبانی، محمدحسین میرقادری	بررسی تحلیلی دیتاست مربوط به شهر نیویورک در سال ۲۰۰۹

۱ چکیده

قصد داریم بر روی یک دیتاست از سال ۲۰۱۹ در مورد Airbnb در نیویورک فرضیاتی مطرح کنیم و به مطالعه آنها بپردازیم. به بررسی تاثیر و رابطه طول جغرافیایی بر قیمت اجاره ها یا بررسی تفاوت قیمتی انواع مسکن های اجاره ای می پردازیم.

کلمات کلیدی

Airbnb ، اجاره، قیمت خانه، نیویورک، طول جغرافیایی.

۲ مقدمه

دیتاست موجود مربوط به تمامی قراداد های اجاره مسکن در سال ۲۰۱۹ هست که توسط پلتفرم Airbnb در شهر نیویورک صورت گرفته است. هر قرارداد شامل اطلاعاتی در مورد اسم مسکن، اسم میزبان، محله ای که مسکن در آن واقع شده است و طول و عرض جغرافیایی مسکن و ... است. در این مورد فرضیاتی مطرح می‌شود که ما در نظر داریم که آن‌ها را بررسی کنیم:

۱. به طور میانگین قیمت اجاره‌ی هر نوع خانه‌ای در محله‌ی منهتن از بقیه محلات بیشتر است

۲. وجود دارد محله ای ک قیمت اجاره‌ی private room ش بیشتر از entire room در محله های دیگر باشد

۳. اول دسته بندی کنیم طول و عرض جغرافیایی رو بعدش برسی کنیم این ادعا رو که بازه ای که بیشترین خونه درون هستند ارزون ترین بازه هم هست.

۴. دلیل تراکم بالای اجاره ها در این بازه $[-74, -73.9]$ طول جغرافیایی، قیمت ارزان تر آن به نسبت دیگر مناطق بوده است

این مسائل می‌توانند از چندین جنبه حائز اهمیت باشند که مهمترین آن‌ها عبارتند از:

- شناسایی خانه مناسب جهت اجاره برای استفاده با در نظر گرفتن نیاز ها و شرایط.

- نرمالیزه کردن قیمت بر اساس دیگر پارامترها و شرایط.
- بررسی پارامتر های تاثیر گذار در قیمت و کیفیت خانه های اجرایی.

۳ روش

۱.۳ بررسی فرضیه‌ی اول

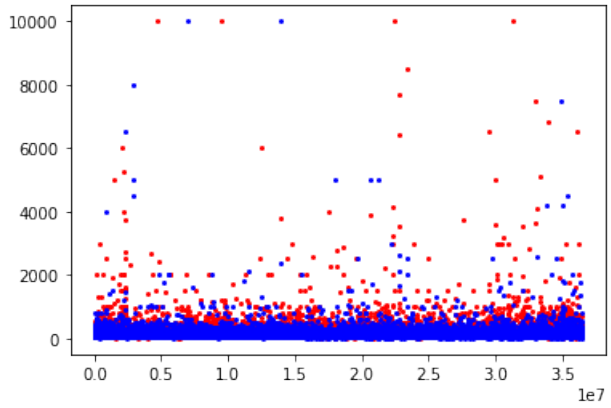
پس از شناسایی داده‌ها متوجه می‌شویم که در این داده‌ها، شهر نیویورک به ۵ گروه محلات تقسیم می‌شوند و در فرضیه‌ی اول ما به بررسی این می‌پردازیم که نشان دهیم به طور میانگین، اجاره‌ی گروه محلات Manhattan از سایر گروه‌ها بیشتر است.

ابتدا ستون‌های موردنیازمان را انتخاب کرده و سایز ستون ها را از آن حذف می‌کنیم. پس از انجام این عملیات و زدن دستور `data.head()` خروجی زیر را خواهیم داشت:

	id	neighbourhood_group	neighbourhood	latitude	longitude	price
0	2539	Brooklyn	Kensington	40.64749	−73.97237	149
1	2595	Manhattan	Midtown	40.75362	−73.98377	225
2	3647	Manhattan	Harlem	40.80902	−73.94190	150
3	3831	Brooklyn	Clinton Hill	40.68514	−73.95976	89
4	5022	Manhattan	East Harlem	40.79851	−73.94399	80

جدول ۱

برای این که بدانیم فرضیه‌مان تا چه حد معتبر است، یک نمودار دوبعدی scatter می‌کشیم. نقاط این نمودار با طول شناسه id و عرض price هستند. نکته‌ای که این نمودار دارد این است که نقاطی که مربوط به خانه‌های گروه محلات Manhattan است به رنگ قرمز و دیگر نقاط به رنگ ابی نشان داده شده است.

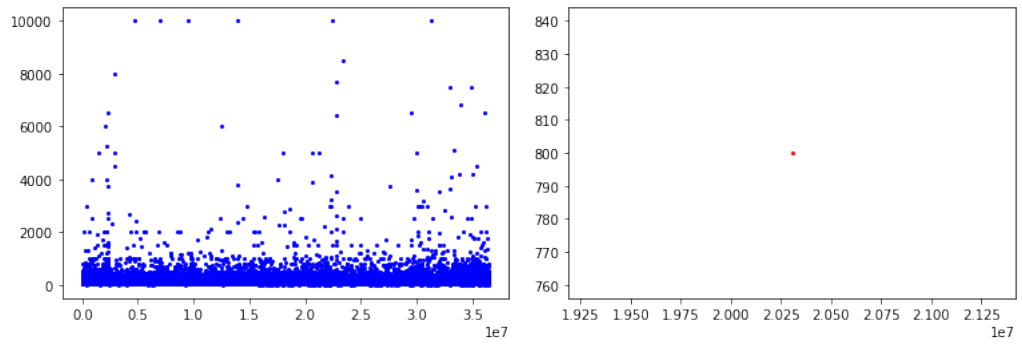


شکل ۱

با در نظر گرفتن این شکل می‌توان نظارت کرد که نقاط قرمز و آبی، تفاوت چندانی با یکدیگر ندارند. یعنی نمی‌توان گفت که قیمت اجاره خانه ها در گروه Manhattan تفاوت زیادی با دیگر گروه ها دارد. پس برای این که فرضیه‌ی خودمان را به اثبات برسانیم، نیاز داریم که یک میانگین گیری ساده روی گروه‌ها و قیمت خانه‌های مربوطه انجام دهیم. برای این کار از کد زیر استفاده می‌کنیم.

```
neighbor_gr_mean_grouped_data = data.groupby(['neighbourhood_group']).mean()
maxprice_neigh_gr = neighbor_gr_mean_grouped_data.idxmax(0)['price']
maxprice_neigh_gr
# OUTPUT: 'Manhattan'
```

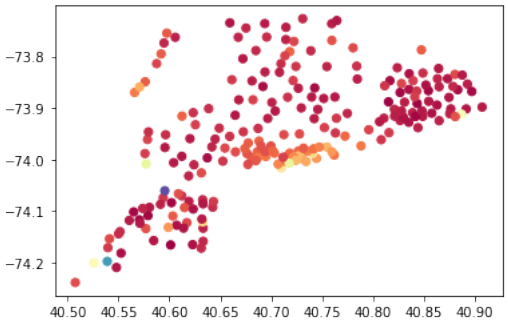
همینطور که کامنت شده است، خروجی این کد همان فرض ماست. در این کد متغیر data در واقع دیتافریم داده‌های ماست. برای بررسی بیشتر نمونه‌ی همین کد را روی محله‌ها به جای گروه‌های محلات اجرا می‌کنیم. پس از این اجرا، خروجی محله‌ی Fort Wadsworth است. مانند شکل ۱، نموداری برای این محله می‌کشیم اما این بار هم بدون خانه‌های سایر محلات و هم با خانه‌های سایر محلات می‌کشیم. نمودار سمت راست، تنها



شکل ۲

شامل یک نقطه است و نشان دهنده‌ی خانه‌ای است که در محله‌ی Fort Wadsworth قرار دارد. یعنی تنها خانه‌ای که در این محله قرار دارد، به طور کلی از میانگین قیمت سایر محلات، قیمت بیشتری دارد.

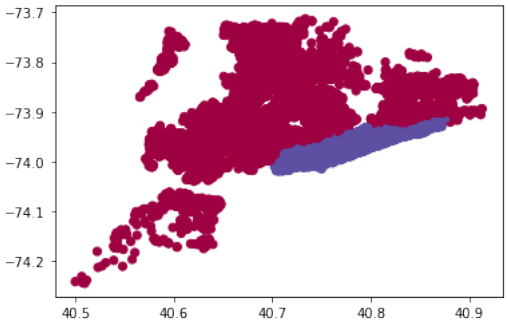
یک راه دیگر برای بررسی میانگین قیمت محلات، استفاده از دو ویژگی طول و عرض جغرافیایی است که با Latitude و Longitude قابل دسترسی هستند. می‌توانیم یک نمودار دوبعدی Scatter رسم کنیم که طول و عرض آن همان طول و عرض جغرافیایی باشتد و علاوه بر آن فیمت ها را با استفاده از یک نگاشت رنگی نمایش دهیم. به صورتی که نقاطی که رنگ سرد تری دارند، کم قیمت تر و نقاطی که رنگ گرم‌تری دارند، قیمت بیشتری دارند. نمودار مربوطه به شکل زیر است:



شکل ۳

برای اینکه به درک بهتری از بررسی ای که می‌خواهیم کنیم برسیم، کافی است، با استفاده از نگاشت رنگی یکسانی، خانه هایی که در گروه Manhattan قرار دارند را به صورت جدا نمایان کنیم.

با مقایسه این دو نمودار مشخص است که خانه‌های گروه Manhattan خانه‌های به نسبت گران‌قیمت تری برای اجاره نسیت به مناطق دیگر هستند، اما این تفاوت قیمت خیلی نیست.

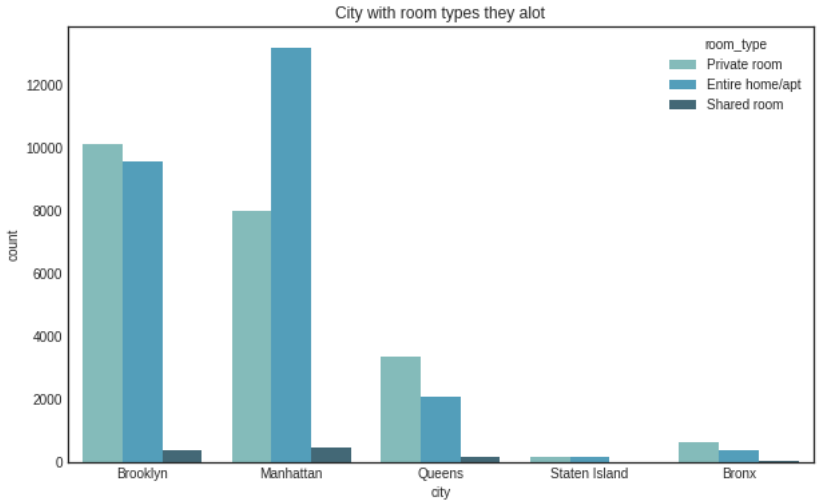


شکل ۴

۲.۳ بررسی فرضیه‌ی دوم

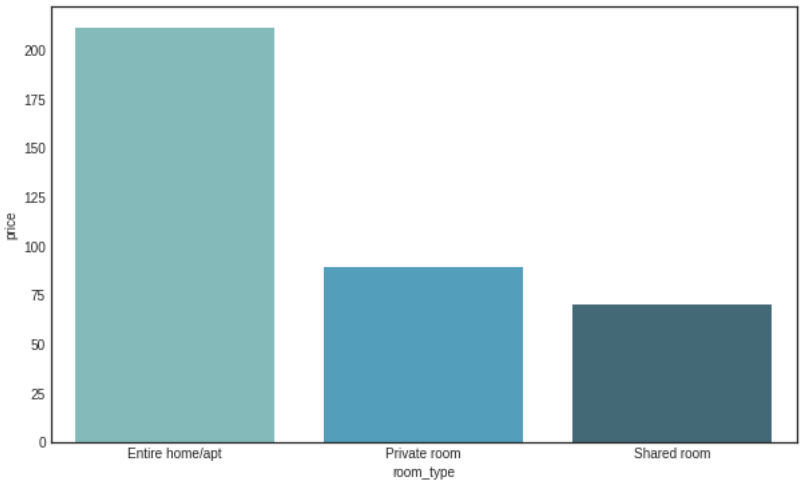
در این فرض بررسی می کنیم که درسترس بودن یک اتاق در طول سال به چه عواملی بستگی دارد. فرض ما این است که دردسترس بودن یک نوع اتاق با تعداد اتاق های موجود از آن نوع رابطه مستقیم و با قیمت آن نوع اتاق رابطه عکس دارد.

برای بررسی عامل اول یعنی تاثیر تعداد اتاق های موجود در هر محله بر دردسترس بودن آن نوع اتاق در طول سال نمودار میله ای زیر را ترسیم می کنیم.



شکل ۵

با توجه به نمودار بالا واضح است که در تمام محله های مورد بررسی اتاق های اشتراکی (shared room) کمترین تعداد اتاق را در مقایسه با دو نوع دیگر در اختیار دارند. در نمودار زیر به بررسی عامل دوم یعنی قیمت انواع اتاق ها می پردازیم.



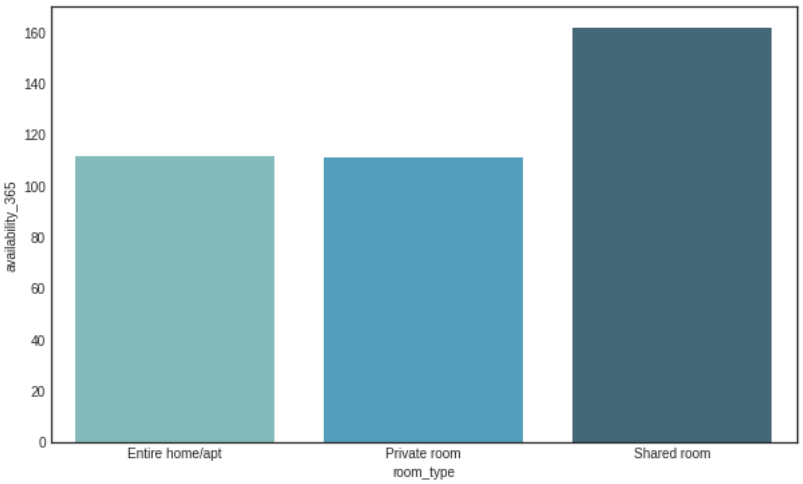
شکل ۶

همانطورکه در نمودار بالا واضح است

۱. قیمت کل خانه بیشتر از سایر انواع است.

۲. اتاق اشتراکی ارزان‌ترین است.

در این مرحله برای مشخص شدن درستی یا نادرستی فرض مان دسترسی انواع اتاق ها در طول سال را در دیتاست خود بررسی می کنیم.

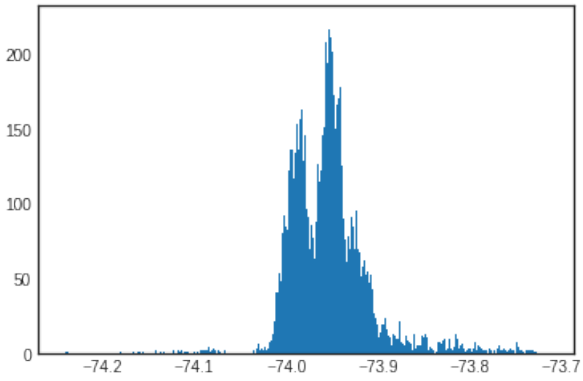


شکل ۷

همانطورکه در نمودار میله ای بالا مشخص است اتاق های اشتراکی در طول سال از سایر اتاق ها بیشتر در دسترس هستند.

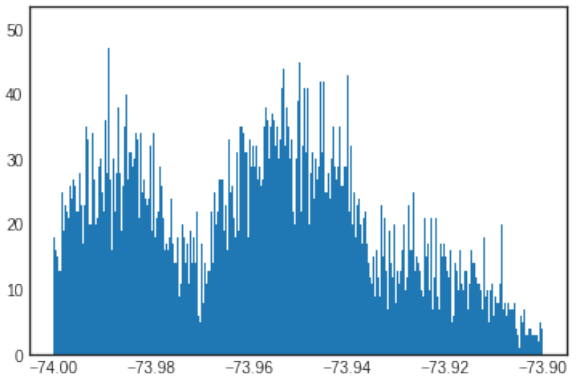
۳.۳ بررسی فرضیه‌ی سوم

فرض این است که بازه طول جغرافیایی که دارای بیشترین تعداد خانه اجاره شده است، میانگین قیمتی پایین تری نسبت به میانگین قیمتی کل خانه های اجاره شده دارد. به منظور تشخیص بازه با بیشترین تعداد خانه اجاره شده، نموندار هیستوگرام مربوط به طول جغرافیایی را میکشیم.



شکل ۸

مشاهده میشود که بیشترین تعداد خانه اجاره داده شده در بازه طول جغرافیایی -73.9 تا -74 قرار دارند. یک دیتاست جدید میسازیم که فقط اطلاعات قرارداد هایی را داشته باشد که طول جغرافیایی آنها در بازه مذکور قرار دارد. حال نمودار هیستوگرام مربوط به سول جغرافیایی را برای داده‌های جدید رسم می‌کنیم.



شکل ۹

نمودار هسیتوگرام دیتاست فیلتر شده را ترسیم میکنیم مشاهده میشود که همه داده ها به درستی در بازه تعیین شده قراردارند. مقدار میانگین قیمت کل قرارداد ها و مقدار میانگین قرارداد هایی که طول جغرافیایی مطلوب را دارند به طور جداگانه حساب میکنیم.

```
total_price_mean = csv['price'].mean()
specific_price_mean = new_df['price'].mean()
print(total_price_mean)
print(specific_price_mean)

# OUTPUT:
# 152.7206871868289
# 151.8987860137149
```

۴.۳ بررسی فرضیه‌ی چهارم

فرضیه این است که دلیل تراکم بالای اجاره ها در این بازه طول جغرافیایی، قیمت ارزان تر آن به نسبت دیگر مناطق بوده است. دو دیتاست از به طور جداگانه از قرارداد های با طول جغرافیایی بیشتر از کران بالای بازه قبلی و از قرارداد های با طول جغرافیایی کمتر از کران پایین بازه قبلی میسازیم. حال مقدار میانگین قیمت قرارداد ها در این دو بازه را محاسبه میکنیم.

```
higher_longitude_df = csv[csv.longitude > -73.9]
lower_longitude_df = csv[csv.longitude < -74]
higher_longitude_mean = higher_longitude_df['price'].mean()
lower_longitude_mean = lower_longitude_df['price'].mean()
print('total price mean:  ' + str(total_price_mean))
print('specific range price:  '  + str(specific_price_mean))
print('higher range price:  ' + str(higher_longitude_mean))
print('lower range price: ' + str(lower_longitude_mean))
# Output:
# total price mean:  152.7206871868289
# specific range price:  151.8987860137149
# higher range price:  93.99702528507684
# lower range price: 228.11330734966592
```

۴ نتیجه‌گیری

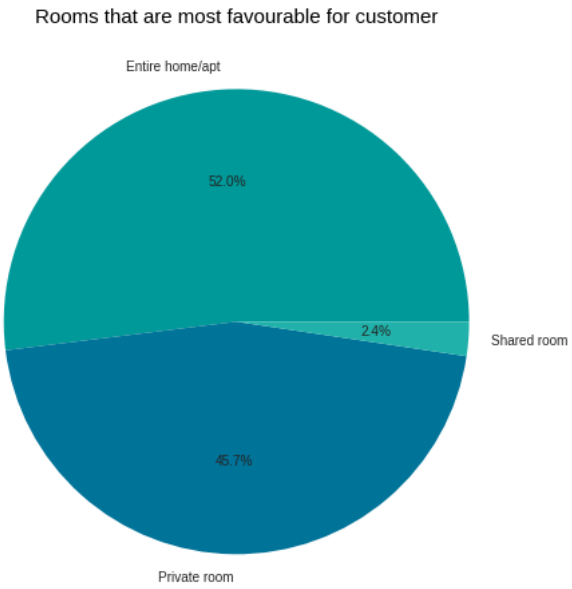
۱.۴ نتیجه‌ی فرضیه‌ی اول

فرض این که قیمت‌های گروه محلات Manhattan بسیار بالا تر از قیمت گروه‌های دیگر است، اشتباه است اما می‌توان گفت در میانگین این فرض درست است. با گرفتن یک میانگین گروهی ساده روی ستون قیمت این نتیجه را میتوان گرفت. در حالت کلی گروه محلات Manhattan از محلات گران است، اما قیمت‌های این مناطق خیلی دور از قیمت های محلات دیگر نیست.

۲.۴ نتیجه‌ی فرضیه‌ی دوم

همانطور که در نمودار های بالا مشاهده کردیم برخلاف فرض، در دسترس بودن یک نوع اتاق به تنهایی وابسته به قیمت و تعداد اتاق موجود نیست. برای نمونه علی رغم اینکه اتاق های اشتراکی کمترین تعداد اتاق را دارا هستند و همچنین کمترین قیمت را در مقایسه با سایر اتاق ها دارند اما در طول سال بیشتر از سایر اتاق ها در دسترس هستند. این نشان دهنده عدم رغبت مسافران به اجاره اتاق های اشتراکی می باشد. به طور کلی می توان گفت که حریم خصوصی ، امکانات رفاهی و مدیریت زمان و فضای سکونت از عوامل و متغیر های اصلی برای انتخاب و اجاره یک اتاق می باشد.

نمودار زیر به وضوح صحت نتیجه گیری بالا را نشان می دهد



شکل ۱۰

۳.۴ نتیجه‌ی فرضیه‌ی سوم

```
total_price_mean = csv['price'].mean()
specific_price_mean = new_df['price'].mean()
print(total_price_mean)
print(specific_price_mean)

# OUTPUT:
# 152.7206871868289
# 151.8987860137149
```

مشاهده میشود که فرضیه ما درست بوده است و میانگین قیمت در این بازه از میانگین قیمت کل قرارداد ها کمی کمتر است. حال میتوان ادعا کرد که دلیل تراکم بالای اجاره ها در این بازه طول جغرافیایی، قیمت ارزان تر آن به نسبت دیگر مناطق بوده است.

۴.۴ نتیجه‌ی فرضیه‌ی چهارم

با توجه به اینکه میانگین قیمت در بازه طول جغرافیایی بزرگتر از کران بالای بازه مطلوب ما، کمتر است از میانگین قیمت در بازه مطلوب، ادعای ما رد میشود چرا که اگر قیمت پایین دلیل تراکم بالای قرارداد ها در آن منطقه بود باید تراکم قرارداد ها در این بازه که میانگین قیمت کمتری دارد بیشتر می‌بود.

۵ منبع

- https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data