به نام خدا

گزارش تمرین 1

محمد سعید حیدری 400422075

مقدمه

در این گزارش اقداماتی که در جریان حل تسک‌های مشخص شده برای تمرین 1 انجام داده شده شرح و توضیحات و تحلیل‌های مربوط به آن‌ها نیز ارائه شده است..

Section 1: Remove missing values and outliers

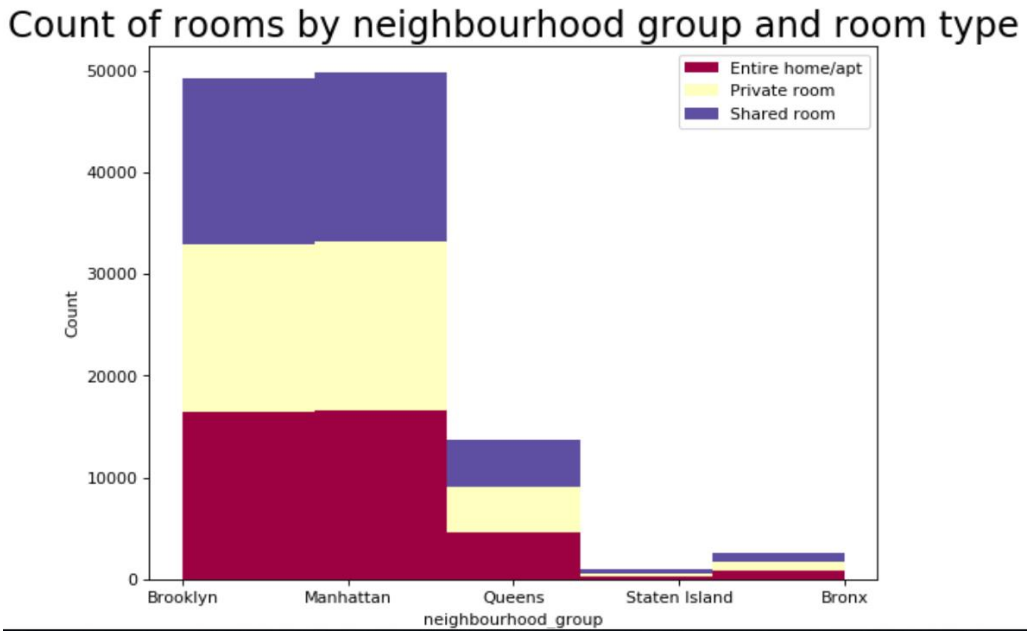
در تسک اول شناسایی و حذف داده‌های میسینگ و داده‌های پرت در دستور کار بوده است. برای این منظور، ما ابتدا با استفاده از دستور pd.dropna() سطرهای حاوی میسینگ را شناسایی و حذف کردیم. سپس، با استفاده از مفهوم نرمال بودن و فاصله نرمالیتی 3 برابری از طرفین میانگین، آن داده‌هایی که فاصله آن‌ها از میانگین کل داده‌ها بیش از 3 انحراف‌معیار فاصله داشت را به عنوان داده پرت شناسایی و حذف کردیم.

$$Outlier \quad if \ (|X_i - Mean|) > 3*Std \tag{1}$$

Section 2: Overal Information

در این تست نمایش اطلاعات کلی مربوط به دیتاست در دستور کار است.

در ابتدا در شکل 1 تعداد خانه‌ها به تفکیک مناطق جغرافیایی و نوع منازل نمایش داده شده‌اند.



شکل 1: فراوانی آگهی‌های منازل به تفکیک مناطق جغرافیایی

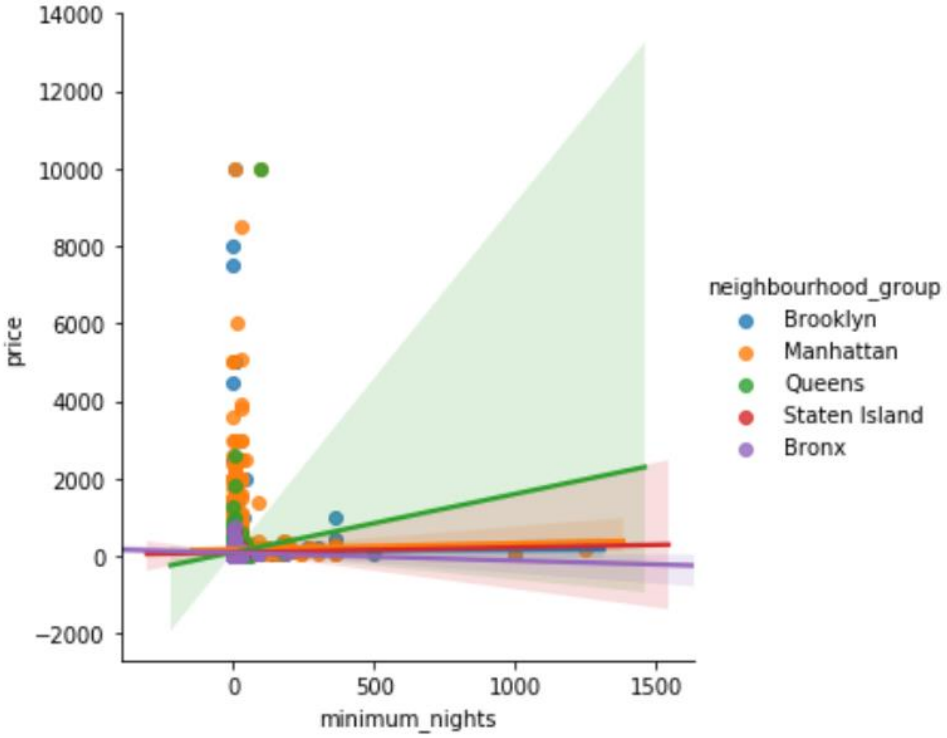
موارد زیر را می توان از این شکل نتیجه گیری کرد:

1) فراوانی آگهی ها در منطقه منهتن بیشترین و بعد از نیز منطقه بروکلین است. در حالی که فراوانی آگهی ها در منطقه استیتن ایسلند و برونکس بسیار پایین تر است.

2) اگر از دید نوع منازل بنگریم، هر یک از سه نوع ممکن منازل سهم برابری فراوانی های یک منطقه دارند. به عبارت دیگر در هر یک از مناطق سهم هر یک از سه نوع مشخص شده از منازل (آپارتمان-اتاق مستقل-اتاق مشترک) یکسان است.

3) تفاوت میان مناطق پر آگهی (منهتن و بروکلین) با مناطق کم آگهی قابل توجه و زیاد است

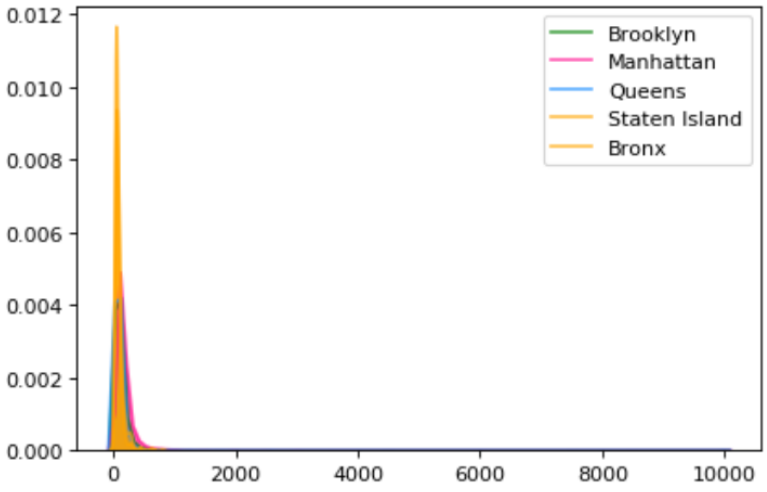
سپس، ما ارتباط میان حداقل شب‌های اقامت را با قیمت به تفکیک هر یک از انواع منازل را در شکل 2 نمایش دادیم.



شکل 2: ارتباط میان حداقل شب‌های قابل اقامت و قیمت

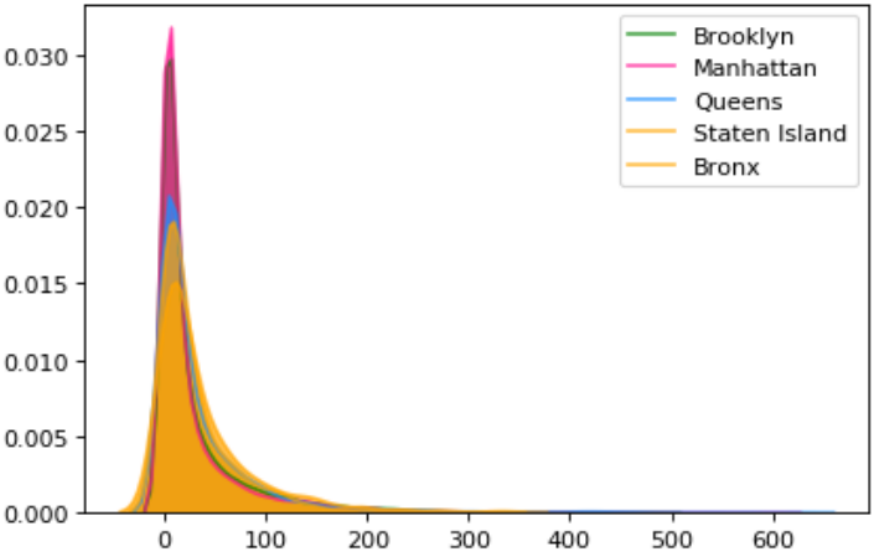
در این شکل تلاش شد تا رابطه میان حداقل نصاب تعداد روزهای اجاره با قیمت در هر یک از سه نوع منزل نمایش داده شود. همانطور که در شکل نیز مشخص است، رابطه مشخص و قابل توجهی بین این دو شاخص در هیچ از یک از سه گروه مختلف منزل ها وجود ندارد. زیرا الگوی پراکندگی بیشتر در یک ناحیه مستقر است و تغییرات قیمت بر اساس حداقل شب‌های اقامت در هیچ یک از مناطق جغرافیایی قابل توجه نیست. تنها در منطقه جغرافیایی Queens اندگی رابطه مثبت دیده می‌شود.

در گام بعدی و در قالب دو شکل زیر تلاش شد تا دیدی کلی از پراکندگی و تمرکز شاخص های قیمت و تعداد نظرات در هر یک از مناطق جغرافیایی ارائه شود.



شکل 3: تابع چگالی قیمت در هر یک از مناطق جغرافیایی

شکل قبل توزیع چگالی قیمت را در هر یک از مناطق جغرافیایی نمایش می دهد. کاملا مشخص است میانگین قیمت در تمام مناطق نزدیک به هم بوده و پراکندگی قسمت بروکلین اندکی تفاوت دارد.

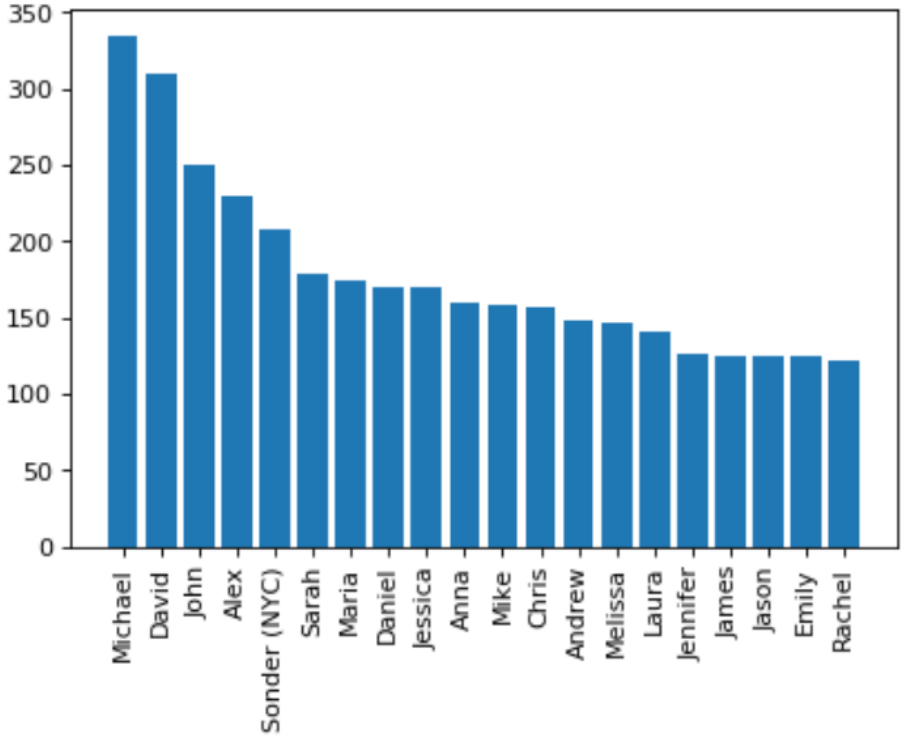


شکل 4: تابع چگالی قیمت در هر یک از مناطق جغرافیایی

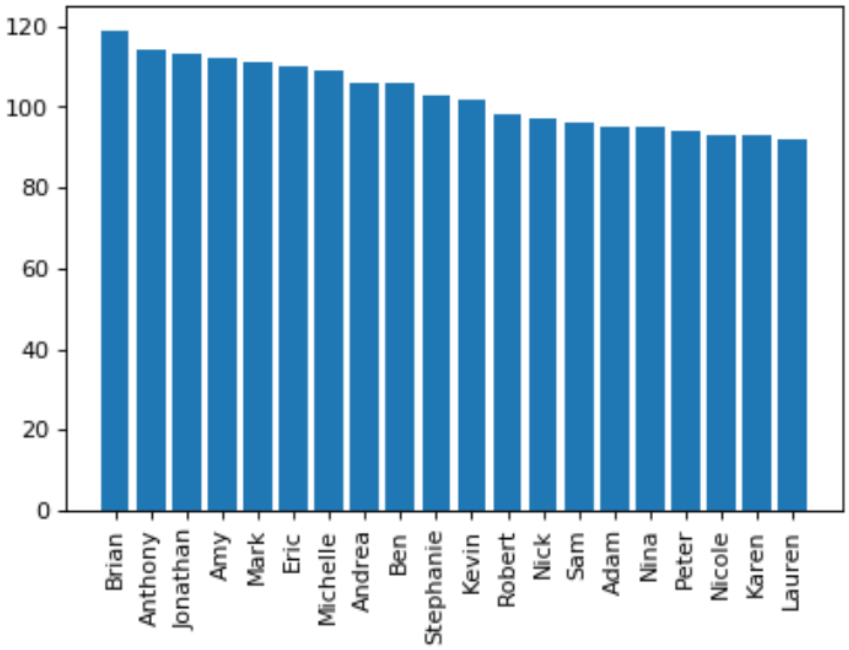
شکل 4 تفاوت در توزیع چگالی تعداد مشتریان را در هر یک از مناطق جغرافیایی نمایش می‌دهد. آنطور که مشخص است، میانگین مشتریان در بروکلین و منهتن بیشتر از بقیه مناطق بوده و از نظر پراکندگی نیز بروکلین بیشترین پراکندگی و منهتن کمترین پراکندگی را دارد.

Section 3: Show announcement owners and their frequencies

در این بخش تعداد نامهای یکتا در صاحبان آگهی شناسایی و نمایش داده می شوند. در این بخش، به دلیل محدودیت نمایش، در دو شکل 5 و 6 به ترتیب فراوانی آگهی های 20 نفر اول و 20 نفر دوم (با بیشترین آگهی ها) نمایش داده می شوند.



شکل 5: فراوانی آگهی‌های مرتبط با 20 نفر اول پرآگهی



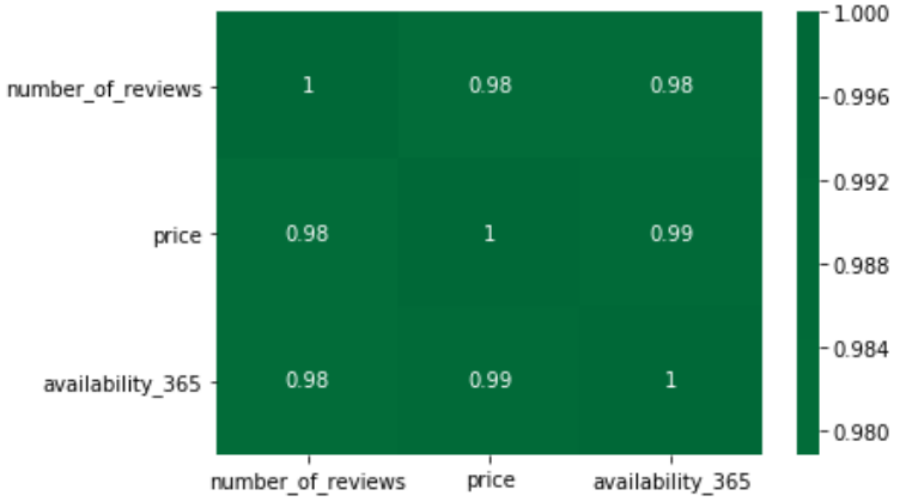
شکل 6: فراوانی آگهی‌های مرتبط با 20 نفر دوم پرآگهی

همچنین، در دو شکل بالا مشخص است که بیشترین فراوانی از آگهی ها با تعداد 335 متعلق به فردی به نام میشل است. طیف فراوانی آگهی ها در میان 40 نفر با بیشترین آگهی ثبت شده از 335 تا 100 متغیر است.

Section 4: Find persons with most costumers and analyze its reasons

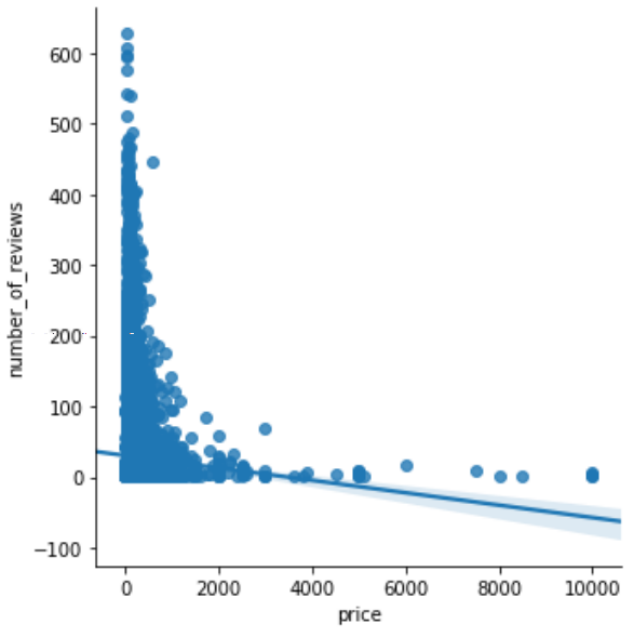
در این قسمت ما از تحلیل همبستگی برای تشخیص میزان ارتباط میان تعداد کامنتهای ثبت شده با دو عامل عددی قیمت و دردسترس بودن استفاده کرده ایم.

طبیعتا، ابتدا نامهای یکتا را شناسایی کرده و سپس نام ها را براساس اعداد آگهی های آن ها سورت کردهایم. سپس، همبستگی میان دو عامل قیمت و در دسترس بودن را در 20 نفر اول از نظر تعداد آگهی ها سنجیدیم. شکل 7 همبستگی میان عوامل قیمت و در دسترس بودن را با تعداد آگهی ها (مشتریان) نمایش داده است.

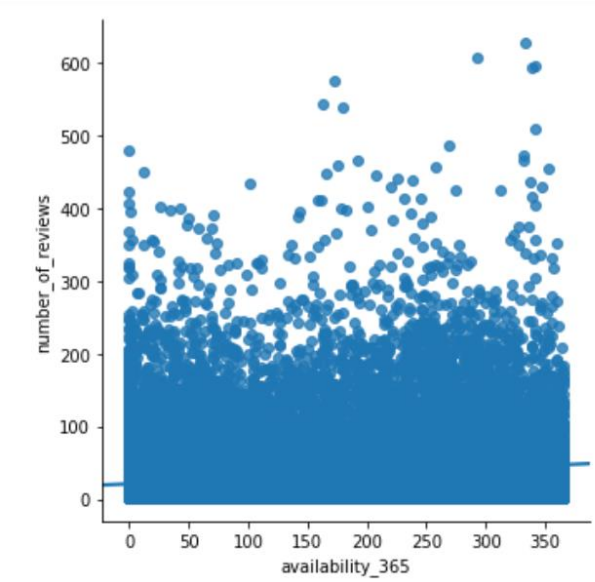


شکل 7: نمایش همبستگی میان دو عامل قیمت و در دسترس بودن و تعداد مشتریان

همانطور که در شکل بالا مشخص است، همبستگی 98 درصدی میان دو عامل قیمت و در دسترس پذیری و تعداد پیامهای ارسال شده به 30 نفر دارای بیشترین مشتری وجود دارد. به عبارت دیگر، همبستگی و ارتباز بالایی میان این عوامل و تعداد مشتری مشتری وجود دارد. برای نمایش نوع این ارتباط از ترسیم دو نمادر اسکتر زیر استفاده شده است. برای تایید این مفهوم، در دو شکل 8 و 9 به ترتیب همبستگی میان قیمت-تعداد مشتریان و همبستگی میان در دسترس‌پذیری –تعداد مشتریان نمایش داده شده است.



شکل 8: نمایش همبستگی میان دو عامل قیمت و تعداد مشتریان



شکل 9: نمایش همبستگی میان دو عامل دسترس پذیری و تعداد مشتریان

به صورت واضح در دو شکل بالا مشاهده می شود که تعداد مشتری با قیمت رابطه معکوس داشته و با دسترس پذیری رابطه مستقیم دارد.

Section 5: Five statistical tests

در این بخش 5 آزمون فرض (4 آزمون یکتا) بر روی مجموعه داده تمرین اعمال شد تا تحلیل‌های درباره نرمال بودن داده‌ها، معنی‌داری روابط میان متغیرها و مواردی از این دست ارائه شود. در ابتدا آزمون فرض نرمال بودن تعداد مشتریان هر آگهی با استفاده از آزمون Shapiro-Wilk انجام شد. در این آزمون فرض صفر و فرض مخالف به صورت زیر هستند:

H0: Number_of_reviews has a Gaussian (normal) distribution

H1: Number_of_reviews does not have a Gaussian (normal) distribution

پس از انجام آزمون، مقدار P_value حاصل 0.0 مشاهده شد که نشانگر عدم تایید فرض صفر بوده و به نظر می‌رسد تعداد پیام‌های هر آگهی از توزیع نرمال پیروس نمی‌کند.

در آزمون فرض دوم، از آزمون همبستگی پیرسون برای سنجش معناداری ارتباط میان تعداد مشتریان و دو عامل قیمت و در دسترسپذیری استفاده کرده ایم. در این آزمون فرض صفر و فرض مخالف به صورت زیر تعریف می‌شوند:

H0: Number_of_reviews and price are independent

H1: There is a dependency between number_of_reviews and price

پس از انجام آزمون، مقدار P-value حاصل از آزمون برابر با 0.0 رویت شد که منجر به نقض فرض صفر می‌شود. به عبارت دیگر، ارتباط میان قیمت و تعداد مشتریان معنی‌دار بوده است.

آزمون فرض سوم نیز مجدد آزمون همبستگی پیرسون اینبار میان دسترس‌پذیری و تعداد مشتریان بوده است. فرض صفر و فرض مخالف به صورت زیر است:

H0: Number_of_reviews and availability_365are independent

H1: There is a dependency between number_of_reviews and availability_365

مقدار P-value حاصل از این آزمون نیز مقدار 0.0 بوده که نشان دهنده نقض فرض صفر و وجود رابطه معنی‌دار میان در دسترس‌پذیری و تعداد مشتریان است.

در آزمون فرض چهارم از آزمون ANOVA برای سنجش همسانی میان قیمت در مناطق مختلف جغرافیایی استفاده کردیم. این آزمون برای سنجش همسانی و تعلق به یک جامعه در میان چند متغیر استفاده می‌شود. فرض صفر و فرض مخالف در این آزمون به صورت زیر هستند:

H0: The means of the samples are equal

H1: One or more of the means of the samples are unequal

پس از انجام آزمون، مقدار P-value برابر با 0.0 حاصل شد که نشان دهنده عدم تایید فرض صفر است. بنابراین، تفاوت معنی‌دار میان قیمت‌ها حداقل در یکی از مناطق تایید می‌شود.

در آزمون فرض آخر نیز با استفاده از آزمون خی دو به بررسی ارتباط بین پرداختیم. در این آزمون فرض وجود ارتباط معنی دار میان دو متغیر با ماهیت غیرعددی منطقه جغرافیایی و نوع اتاق با متغیر هدف تعداد مشتریان سنجیده می شود. در ابتدا ارتباط میان متغیر منطقه جغرافیایی و تعداد مشتریان سنجیده شد. در این آزمون فرض صفر و فرض مخالف به صورت زیر هستند:

H0: Neighbourhood_group and number_of_reviews are independent

H1: There is a dependency between Neighbourhood_group and room_type

با توجه به P-value حاصله فرض صفر آزمون بالا تایید نشده و در نتیجه وابستگی ای میان دو متغیر منطقه جغرافیایی و تعداد مشتریان وجود دارد. به عبارت دیگر، منطقه جغرافیایی در تعداد مشتریان تاثیرگذار است. در ادامه، تلاش شد تا ارتباط میان متغیر کیفی نوع اتاق با تعداد مشتریان نیز سنجیده شود. در این آزمون فرض صفر و فرض مخالف به صورت زیر هستند:

H0: room_type and number_of_reviews are independent

H1: There is a dependency between Neighbourhood_group and room_type

با توجه محاسبه pvalue=0 فرض صفر آزمون بالا نیز تایید نشده و در نتیجه وابستگی میان دو متغیر نوع اتاق و تعداد مشتریان وجود دارد. به عبارت دیگر، نوع اتاق یا منزل نیز در تعداد مشتریان تاثیرگذار است.

Section 6: Forecast price and number of reviews

در این بخش ما با استفاده از دو مدل رگرسیون به ترتیب قیمت و تعداد مشتریان را از روی دیگر متغیرها پیش‌بینی کردیم. در این گام، سعی شد تا پس از ساختن مدل رگرسیونی، ضرایب رگرسیونی برای مشاهده میزان اهمیت هر یک از متغرهای ورودی تحلیل شوند. در این راستا، ابتدا مقادیر متنی را به صورت کدهای عددی تبدیل کردیم. برای ساخت دو مدل رگرسیونی نیز از کتابخانه scikit learn استفاده کردیم.

ابتدا مدل رگرسیونی پیش‌بینی قیمت در دستور کار قرار گرفت. پس از ساخت مدل رگرسیونی اول، ضرایب متناظر به هر یک از ورودی ها به صورت شکل 10 محاسبه شدند.

```
Out[21]: Pipeline(steps=[('transformedtargetregressor',
                          TransformedTargetRegressor(func=<ufunc 'log10'>,
                                                     inverse_func=<ufunc 'exp10'>,
                                                     regressor=Ridge(alpha=1e-1
         0)))])
```

```
In [13]: print(X.columns)

         Index(['neighbourhood_group', 'latitude', 'room_type', 'minimum_nights',
                'calculated_host_listings_count', 'availability_365'],
               dtype='object')
```

```
In [12]: print(MLR.coef_)

         [-1.17123240e+01  2.27862531e+02  8.51171786e+01 -7.49887039e-03
           1.79360053e-01  1.21959807e-01]
```

شکل 10: ضرایب هر یک از متغیرهای ورودی در مدل رگرسیونی آموزش دیده جهت پیش‌بینی قیمت

همانطور که در شکل 10 مشخص است، از میان متغیرها مستقل موجود در دیتاست، latitude، room_type و neighbourhood_group دارای بیشترین تاثیر و اهمیت در ساختن مدل پیش بینی قیمت بوده اند.

پس از آن، ساخت مدل پیش‌بینی تعداد مشتریان دنبال شد. شکل 11 ضرایب نهایی مدل آموزش دیده برای پیش‌بینی تعداد مشتریان را از روی متغیرهای مستقل مجموعه داده نمایش می‌دهد.

ساخت مدل رگرسیونی برای پیش بینی تعداد مشتری:

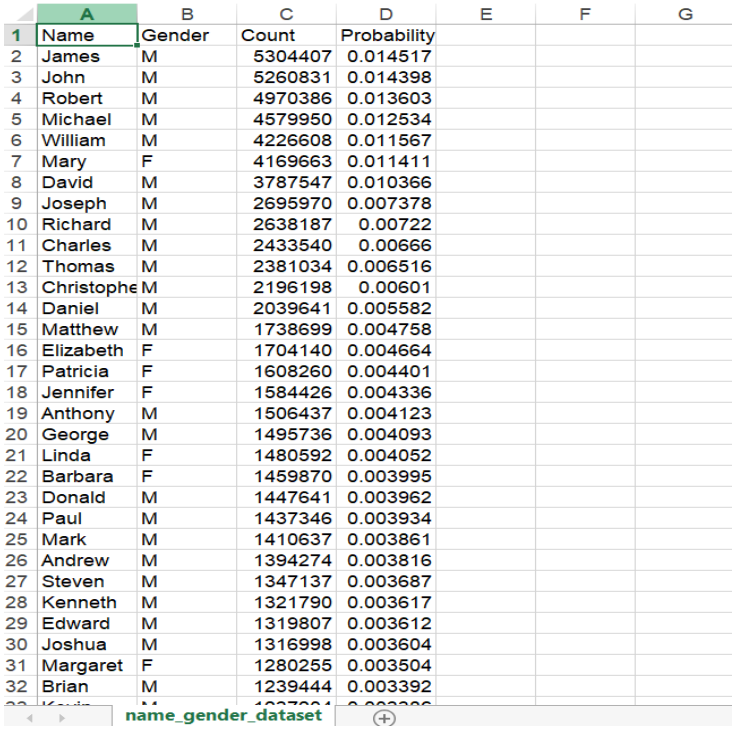
```
In [14]: MLR = LinearRegression().fit(X, Y2)

In [15]: print(MLR.coef_)
         print(X.columns)

[-0.30977897  1.24290268 -1.44087565 -0.23274646 -0.16951436  0.08170939]
Index(['neighbourhood_group', 'latitude', 'room_type', 'minimum_nights',
       'calculated_host_listings_count', 'availability_365'],
      dtype='object')
```

شکل 11: ضرایب هر یک از متغیرهای ورودی در مدل رگرسیونی آموزش دیده جهت پیش‌بینی تعداد مشتری

همانطور که در شکل 11 مشخص است، از میان متغیرها مستقل موجود در دیتاست، latitude، room_type و calculated_host_listings_count_group دارای بیشترین تاثیر و اهمیت در ساختن مدل پیش بینی تعداد مشتریان بوده اند.

Section 7: Analyze relation between gender and price/costumer

در این بخش ما ابتدا یک مجموعه داده تهیه شده از گیت‌هاب حاوی اسامی انگلیسی و جنسیت آن‌ها را فراخوانی کردیم (شکل 12).



	A	B	C	D	E	F	G
1	Name	Gender	Count	Probability			
2	James	M	5304407	0.014517			
3	John	M	5260831	0.014398			
4	Robert	M	4970386	0.013603			
5	Michael	M	4579950	0.012534			
6	William	M	4226608	0.011567			
7	Mary	F	4169663	0.011411			
8	David	M	3787547	0.010366			
9	Joseph	M	2695970	0.007378			
10	Richard	M	2638187	0.00722			
11	Charles	M	2433540	0.00666			
12	Thomas	M	2381034	0.006516			
13	Christophe	M	2196198	0.00601			
14	Daniel	M	2039641	0.005582			
15	Matthew	M	1738699	0.004758			
16	Elizabeth	F	1704140	0.004664			
17	Patricia	F	1608260	0.004401			
18	Jennifer	F	1584426	0.004336			
19	Anthony	M	1506437	0.004123			
20	George	M	1495736	0.004093			
21	Linda	F	1480592	0.004052			
22	Barbara	F	1459870	0.003995			
23	Donald	M	1447641	0.003962			
24	Paul	M	1437346	0.003934			
25	Mark	M	1410637	0.003861			
26	Andrew	M	1394274	0.003816			
27	Steven	M	1347137	0.003687			
28	Kenneth	M	1321790	0.003617			
29	Edward	M	1319807	0.003612			
30	Joshua	M	1316998	0.003604			
31	Margaret	F	1280255	0.003504			
32	Brian	M	1239444	0.003392			

name_gender_dataset

شکل 12: نمایشی از مجموعه داده اسامی و جنسیت

سپس، از آزمون خی دو برای سنجش ارتباط میان جنسیت و دو عامل قیمت و تعداد مشتریان بهره گرفتیم. فرض صفر و فرض مخالف در این دو آزمون به صورت زیر بوده است:

H0: Sex and price/number_of_reviews are independent

H1: There is a dependency between Sex and price/number_of_reviews are independent

پس از انجام دو آزمون، مقدار P-value حاصل از هر دو 0.23 و 0.37 محاسبه شد. در نتیجه به نظر می‌رسد رابطه معنی‌داری میان جنسیت و دو عامل قیمت و تعداد مشتریان وجود ندارد.

نتیجه گیری کلی : تمرکز این پروژه بیشتر روی تحلیل روابط بین متغییر هاست و سعی میکنه انواع تحلیل های بصری و منطقی رو تهیه کنه . متغییر هایی مثل قیمت و نوع خانه و موقعیت توی تعداد مشتری تاثیر دارن. گزارش از تمامی مراحل به صورت استپ بای استپ نیز خدمتتان ارسال شده است

جناب شریفی

پیروز و سرافراز باشید