پروژه درس داده‌کاوی

دانشگاه شهید بهشتی

دانشکده علوم ریاضی

حمیدرضا فیروزه(400422147)

استاد: دکتر هادی فراهانی

دستیار استاد: مهندس علی شریفی

بهار 1401

فهرست مطالب

1- مقدمه

هدف این پروژه بررسی دو دیتاست ارائه شده در سایت Kaggle است. دیتاست اول اطلاعات واقعی در مورد خانه های به اشتراک گذاشته در سایت Airbnb در شهر نیویورک است. دیتاست دوم هم اطلاعات مرتبط با ملک‌های ارائه شده جهت اجاره در آلمان است.

به منظور بررسی این داده‌ها از Google Colab استفاده شده است. برای هر دیتاست یک سری مراحل به ترتیب اجرا شده است. در ابتدا آنالیز اکتشافی داده‌ها(Exploratory Data Analysis-EDA) انجام گرفته شده است. سپس از روشهای یادگیری ماشین برای پیش بینی پارامترهای مدنظر براساس فیچرهای در دسترس استفاده شده‌است. علاوه بر این برای هر دیتاست تسک‌های از پیش تعریف شده‌ای وجود دارد که، سعی خواهیم کرد تا این تسک‌ها به همراه تحلیل‌های لازم را انجام دهیم. نکته قابل توجه این است که در این گزارش کدها ثبت نشده است و فقط خروجی ها و تحلیل آنها آورده شد.

2- دیتاست یک

این مجموعه داده‌ها شامل داده‌های واقعی از خانه‌هایی است که توسط میزبانان در سایت Airbnb در نیویورک به اشتراک گذاشته شده‌است. اهداف اصلی این پروژه عبارتند از:

1- پاکسازی داده از قبیل بررسی داده های از غیرموجود و یافتن داده های پرت.
2- ارائه اطلاعات کلی در حالت تجمیعی در خصوص اگهی از قبیل تعداد آ گهی ها ، تعداد آگهی ها در هر منطقه جغرافیایی ، بررسی شاخص های کلی قیمت و ... .
3- بررسی صاحبان آ گهی و گزارشی از تعداد خانه های مرتبط با هر صاحب آ گهی
4- اگر تعداد کامنت های برای یک آ گهی را بتوان شاخصی از تعداد مشتریان در نظر گرفت مطلوب است یافتن صاحبان آ گهی که بیشترین مشتری را دارا می باشند و بررسی علت های آن.
5- مطرح کردن ۵ آزمون فرض دلخواه در داده ها و پاسخ گویی و تفسیر آنها (حداقل از ۳ آزمون فرض متفاوت استفاده کنید.)
6- تلاش در ساخت مدل برای پیش بینی پارامترهایی همانند قیمت و ارایه این مدل ها و تفسیر آنها

تسک های امتیازی

۱ -اضافه کردن اطلاعات اضافی به دیتاست از قبیل اطلاعات جغرافیایی ایستگاه های مترو و موزه ها ، فرودگاه ، ایستگاه قطار مرکزی و ... و بررسی اثر گذاری این شاخص ها بر آ گهی از قبیل قیمت و تعداد مشتری و میزان رضایتمندی مشتریان و ...

۲- بررسی نقش زن بودن یا مرد بودن صاحب آ گهی در قیمت و تعداد مشتریان و میزان رضایتمندی مشتریان

2-1- پاکسازی داده ها

بعد از خواندان داده ها از فایل csv، این داده ها در دیتافریمی به اسم raw_data ذخیره شده است. در ابتدا یک نگاه اجمالی به داده های موجود خواهیم داشت.

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	nu
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	

شکل 1بخشی از داده های موجود در دیتاست

تعداد کل داده های موجود برابر با 48895 رکورد با 16 ستون بوده است.

نوع داده ها مجموعه ای از داده های عدد صحیح، عدد حقیقی و آبجکت است که در شکل زیر قابل مشاهده است.

```
id                                int64
name                             object
host_id                           int64
host_name                        object
neighbourhood_group              object
neighbourhood                    object
latitude                        float64
longitude                       float64
room_type                        object
price                             int64
minimum_nights                    int64
number_of_reviews                 int64
last_review                      object
reviews_per_month               float64
calculated_host_listings_count    int64
availability_365                  int64
dtype: object
```

شکل2 ماهیت داده های موجود با استفاده از دستور dtypes

2-1-1 بررسی داده‌های گمشده

به منظور پاکسازی داده ها ابتدا بررسی می کنیم که چه تعداد داده گمشده در هر ستون وجود دارد.

```
id                                    0
name                                 16
host_id                               0
host_name                            21
neighbourhood_group                   0
neighbourhood                         0
latitude                              0
longitude                             0
room_type                             0
price                                 0
minimum_nights                        0
number_of_reviews                     0
last_review                       10052
reviews_per_month                 10052
calculated_host_listings_count        0
availability_365                      0
dtype: int64
```

شکل 3:تعداد داده های گمشده برای هر فیچر

name و host_name دارای ۱۶ و ۲۱ مقدار گم‌شده می‌باشد. بر اساس لیست تسک های مرتبط با این دیتاست ما از این ویژگی name استفاده نخواهیم کرد بنابراین می‌توانیم آن را حذف کنیم. با این حال برای ستون نام میزبان ممکن است نیاز به پیش‌بینی جنسبت مالک داشته باشیم، پس این ستون رو حفظ می کنیم و در تسک امتیازی مربوطه داده های گمشده مربوط به این ستون را بررسی می کنیم. خروجی ما شان می‌دهد که داده‌های گم‌شده برای last_review و review_per_month زمانی اتفاق می‌افتد که هیچ نظری برای این ملک ثبت نشده باشد برای بررسی این موضوع که این دو ستون را تجزیه و تحلیل می‌کنیم. ابتدا حداقل مقدار review_per_month رو چک می کنیم که برابر با 0.1 است. درنتیجه ما عدد صفر نداریم و همانطور که می‌بینیم فرض ما صحیح است و مقادیر از دست رفته برای این دو ستون زمانی اتفاق می‌افتد که هیچ بررسی برای این ویژگی وجود ندارد.

در اینصورت ما می‌توانیم مقادیر Nan در ستون review_per_month را با صفر در هر ماه عوض کنیم و ۲۰۲۲ - ۰۱ - ۰۱ را برای last_review اختصاص دهیم که نشان می‌دهد این ویژگی ممکن است در آینده مورد بازبینی قرار گیرد.

2-1-2- بررسی داده‌های پرت

- **بررسی neighbourhood_group و neighbourhood**

برای بررسی این دو ستون از متد value_counts استفاده می‌کنیم.

```
Manhattan        44.301053
Brooklyn         41.116679
Queens           11.588097
Bronx             2.231312
Staten Island     0.762859
Name: neighbourhood_group, dtype: float64
```

شکل 4 تعداد داده های برای هر گروه در ستون neighbourhood_group

به نظر می‌رسد که منطقه Staten Island فقط 0.76 درصد کل داده‌ها را به خود اختصاص داده‌است. ما می‌توانیم بخشی از داده‌ها را حذف کنیم. با این حال ترجیح می‌دهم داده‌های این منطقه حفظ شوند تا تاثیر محله قیمت را بررسی کنند.

```
Williamsburg          3920
Bedford-Stuyvesant    3714
Harlem                2658
Bushwick              2465
Upper West Side       1971
                      ...
Fort Wadsworth           1
Richmondtown             1
New Dorp                 1
Rossville                1
Willowbrook              1
Name: neighbourhood, Length: 221, dtype: int64
```

شکل 5 تعداد داده ها برای هر گروه در ستون neighbourhood

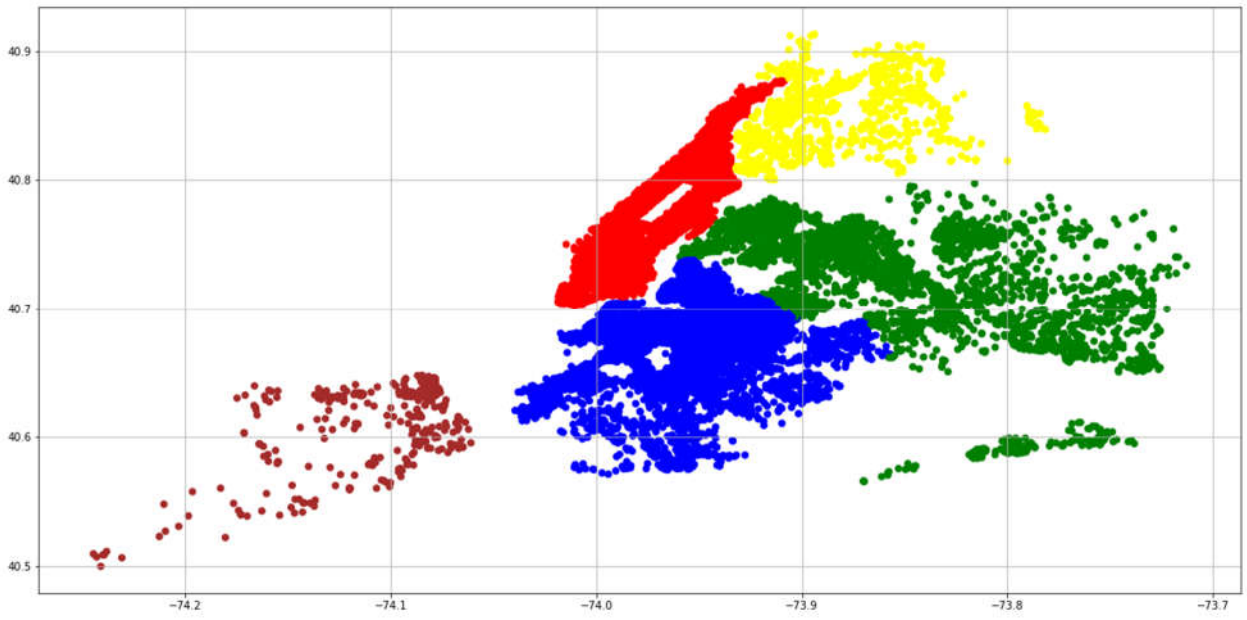
توزیع مقادیر neighberhood نشان می‌دهد که ۲۲۱ محله مجزا وجود دارد. با این حال، برخی محله در فرکانس پایین وجود دارند که تعداد آنها هم قابل توجه است. درنتیجه منطقی نیست که آن‌ها را از مجموعه داده‌ها حذف کنیم.

- **تحلیل طول و عرض جغرافیایی**

برای بررسی تقاط دورافتاده بر اساس طول و عرض جعرافیایی، نمودار پراکندگی ابزار مناسبی خواهد بود.

در نمودار پراکندگی که براساس هر منطقه در نیویورک رنگ هر نقطه مشخص شده. در این نمودار پراکندگی نقطه ای دیده نمیشود که خیلی خارج از محدوده نیویورک باشد. در واقع براساس نمودار زیر، می‌توانیم ببینیم که برخی از خانه‌ها د Staten Island(قهوه‌ای)، Queens (سبز)و Bronx (زرد)به گونه‌ای از بخش‌های دیگر شهر جدا شده‌اند.

با این حال، ترجیح بر این است که این نقاط حذف نشوند تا بعدا بررسی شود که آیا حضور در این نقاط تاثیری بر روی داده های عدید مثل قیمت دارد یا نه؟



شکل 6نمودار پراکندگی براساس طول و عرض جغرافیایی و منطقه

- **بررسی نوع خانه‌ها**

ابتدا تعداد انواع خانه ها را بررسی می کنیم.

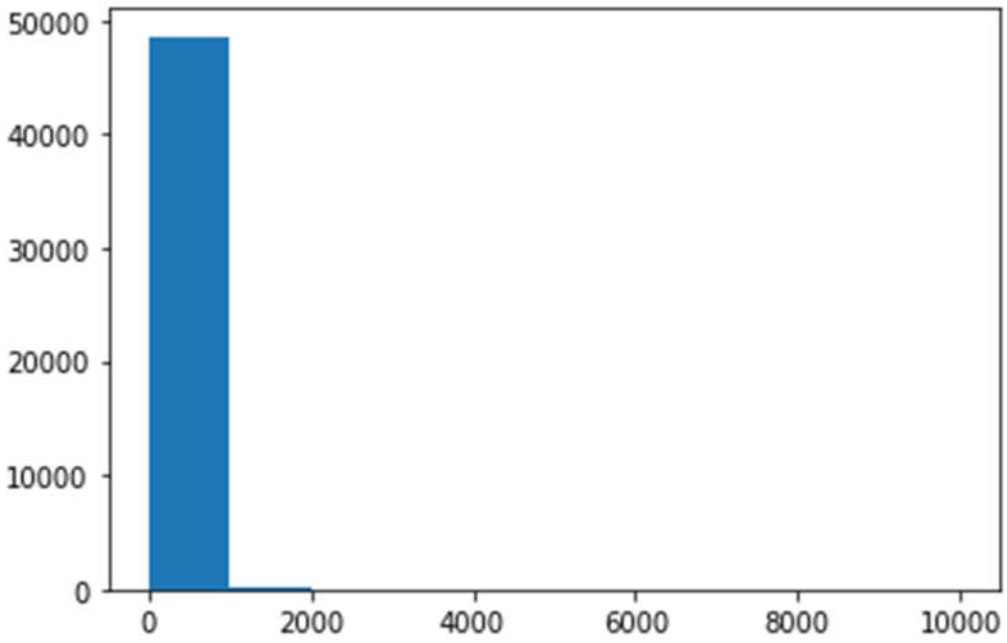
```
Entire home/apt    25409
Private room       22326
Shared room         1160
Name: room_type, dtype: int64
```

شکل 7 فرکانس انواع خانه ها

اینجا من داده ای نمی بینم که دورافتاده باشد

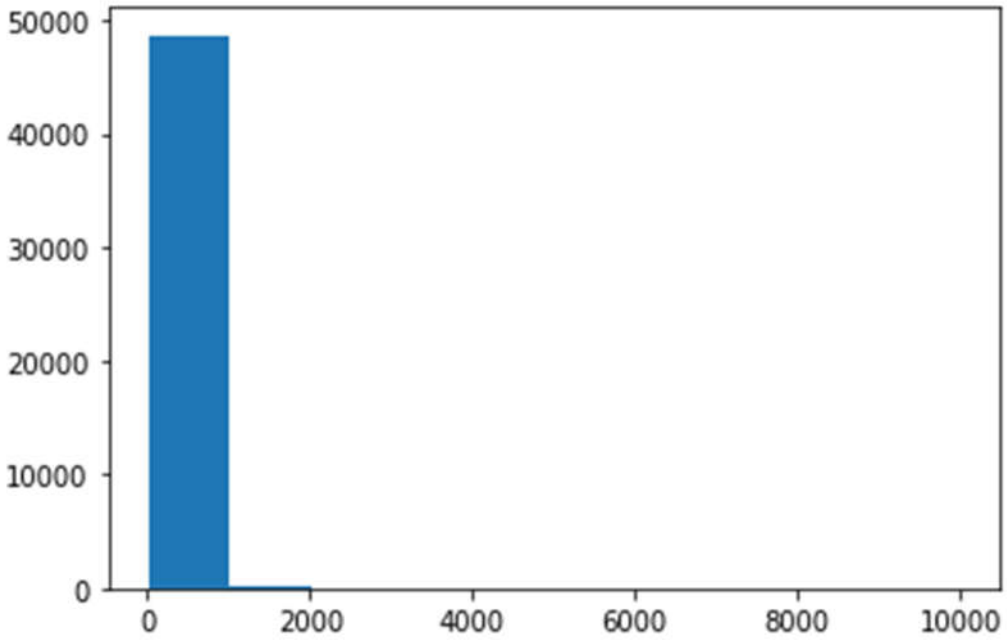
- **بررسی قیمت**

ابتدا یک هیستوگرام براساس قیمت رسم می کنیم. بر اساس هیستوگرام زیر نمی توان هیچگونه تحلیل خاصی نمی توان کرد.

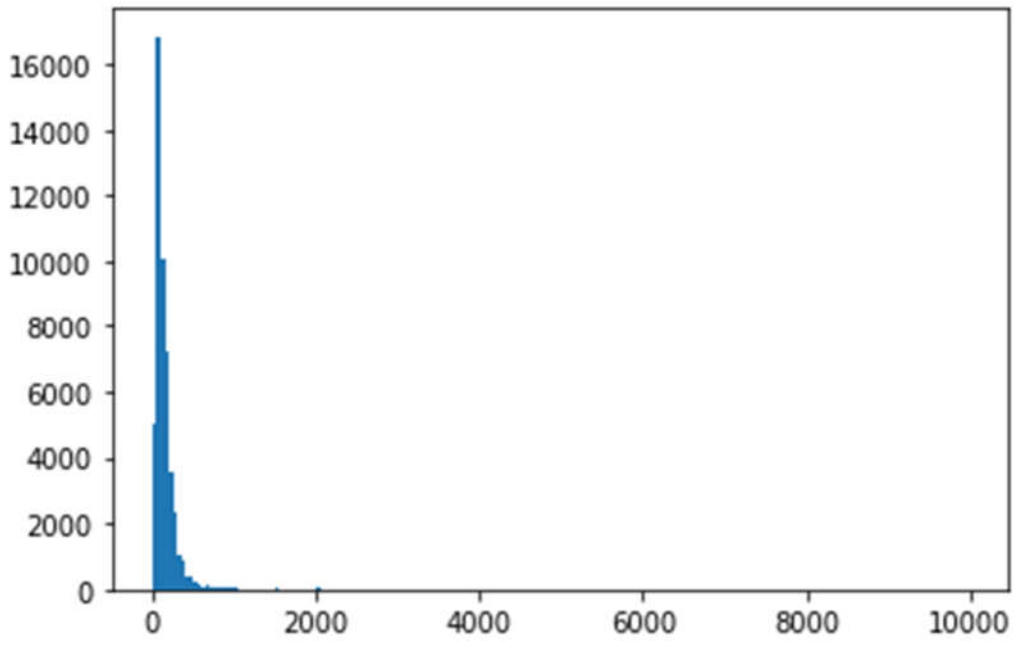


قیمت حداقل خانه ها برابر صفر است که غیر منطقی است، و قیمت حداکثر خانه ها 1000 است.

ابتدا خانه هایی که قیمتشان صفر خورده را حذف می کنیم و دوباره هیستوگرام را رسم می کنیم.



همچنان از روی نمودار بالا نمی توان تحلیل خاصی کرد. به همین دلیل انداز هر بین رو از پیش تعریف می کنیم و دوباره هیستوگرام را رسم می کنیم. بازه های از 0 تا10000 و اندازه هر بین 50 است.



نمودار کمی بهتر شده است ولی هنوز قابل استفاده نیست.

با این حال می توان دید که این که داده‌ها از توزیع نرمال برخوردار هستند. من ترجیح می‌دهم از روش IQR استفاده کنم. باید در نظرداشت که از حد بالا برای حذف outliers استفاده نمی‌کنم.

استفاده از IQR باعث حذف 6 درصد از داده ها شد که مقدار معقولی است. توضحیات مربوط به روش IQR را در اینجا می توانید مشاهده کنید.

- **Minimum nights**

اعداد بیش از یک سال ممکن است غیرمعمول به نظر برسد. مالک ممکن است به صورت تصادفی عددی وارد کرده باشد. بنابراین بهتر است که به رکوردهای بیش از ۳۶۵ روز نگاهی بیاندازیم.

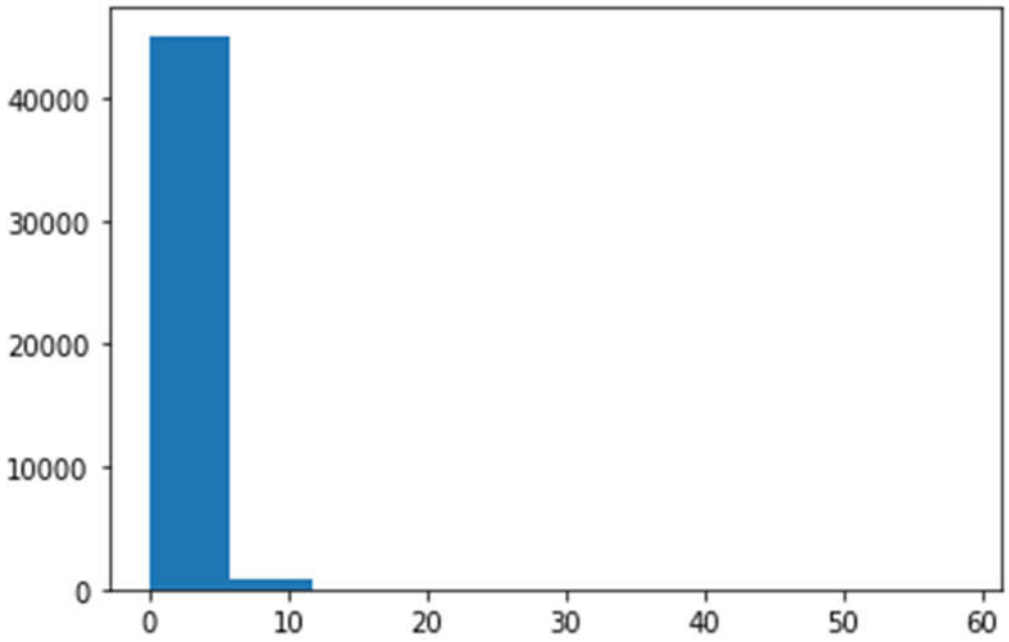
تعداد این داده های خیلی زیاد به نظر نمی رسد. بنابراین این داده ها را با عدد 365 جایگزین می کنیم.

- تعداد کامنت ها به ازای هر ماه

ابتدا هیستوگرام این ستون را رسم می کنیم.

همانطور که از شکل پایین مشخص است برداشت خاصی نمی توان از این داده‌ها داشت.

به همین دلیل به رکوردهایی نگاه می کنیم که تعداد کامنت‌های آنها بیش از 15 عدد در ماه هستند. به طور کلی داشتن این تعداد کامنت در ماه به معنی این است که 15 نفر یا بیشتر مکان میزبان را اجاره کرده اند هر کدام حداکثر دو روز رزرو داشته اند. به نظر من این عدد غیر منطقی است.



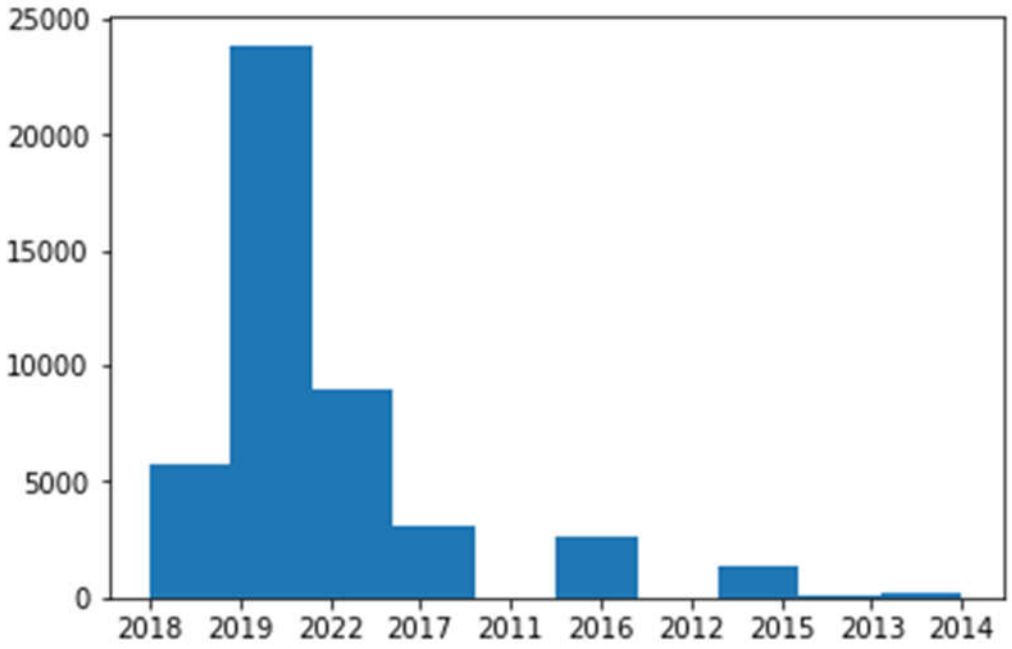
همانطور که در نمودار پایین می‌بینیم در هر ماه تعداد کمی موارد با بیش از ۱۵ مورد بررسی وجود دارد. به نظر من، داشتن بیش از ۱۵ مورد کامنت در هر ماه دشوار خواهد بود. مگر اینکه، کامنت جعلی باشد.

با این حال، بهتر است این داده ها حذف نشود. چون بررسی تاثیر انها بر قیمت می تواند جالب باشد.

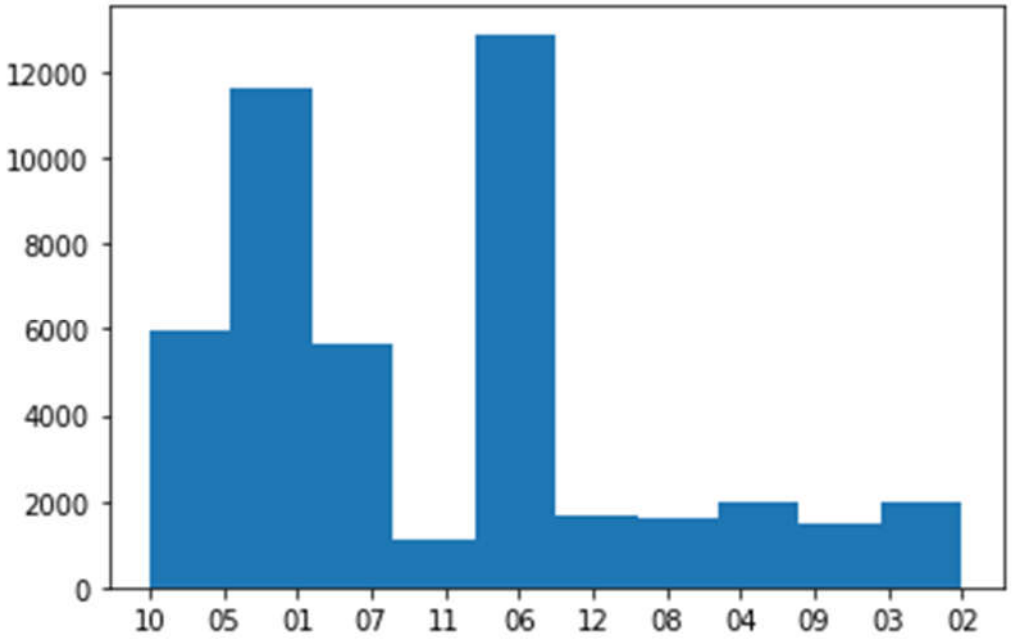
- آخرین کامنت(last_review)

به منظور بررسی این مورد، داده های این بخش به دو ستون last_review_year و last_review_month تبدیل می شوند. به این ترتیب ما میتوانی اثر این دو مورد را باهم بررسی کنیم.

ابتدا تاثیر سال را بررسی می کنیم. هیستوگرامی بر اساس تعداد دفعات تکرار هر سال برای اخریت کامنت رسم می‌کنیم.



شکل بالا منطقی به نظر می رسد. از این شکل، ما می‌توانیم فرض کنیم که برخی موارد بوده اند که دیگر کسی برای آنها کامنتی نگذاشته که می تواند به دلیل خروج این خانه از لیست موارد قابل رزرو باشد. با این حال، نرخ این خانه‌ها در حال افزایش است. به نظر من، این منطقی است، چون شرکت رشد کرده و خانه‌های بیشتر و بیشتری به اشتراک گذاشته می‌شوند و بنابراین خانه‌های بیشتری از فرآیند به اشتراک گذاری خارج می‌شوند.

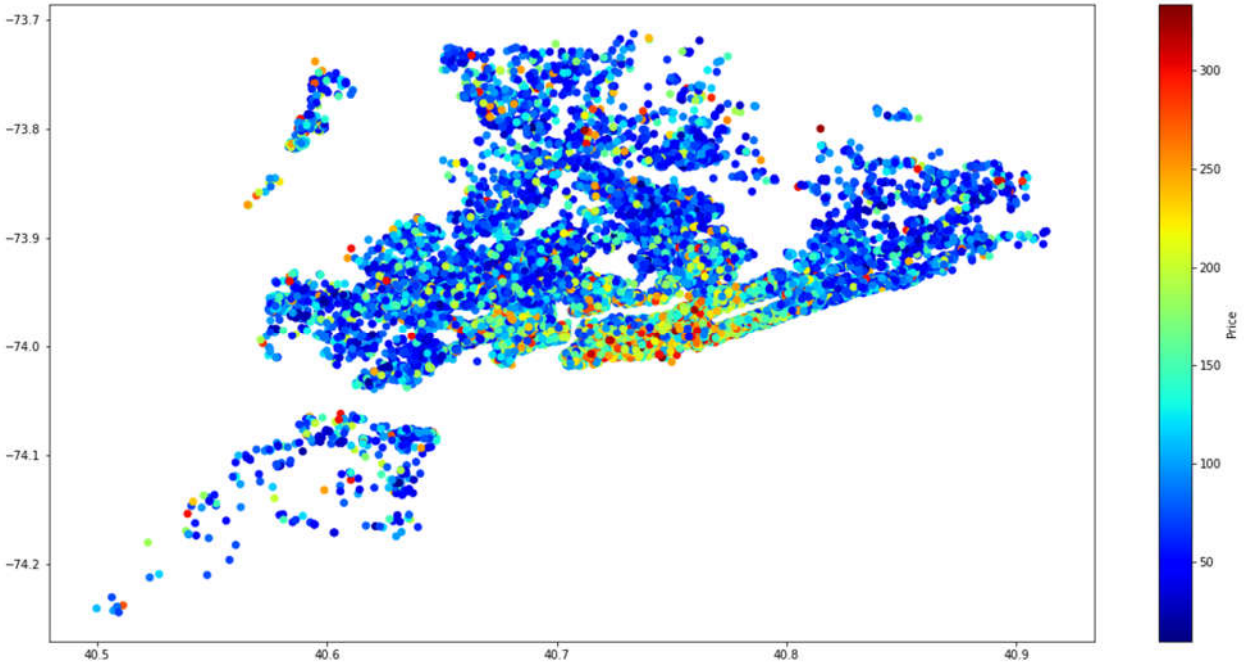


توزیع ماه کمی غیر منطقی به نظر میرسد. بیشتر انتظار توزیع یکنواخت برای این مورد میرفت. افزایش فراوانی در ماه ۰۱، منطقی است. ما مقدار Nan را با ۲۰۲۲ - ۰۱ - ۰۱ جایگزین کردیم. بنابراین ما همه مقادیر گمشده را به عنوان ۰۱ داریم.

با این حال، من نمی‌توانم نتیجه‌گیری کنم که چرا آخرین کامنت برای بسیاری از خانه ها در ماه ژوئن انجام شد؟؟ شاید به دلیل اینکه این داده ها در ابتدای ماه جولای گرفته شده و اکثر خانه ها در ماه قبل کامنت داشتند.

2-2 تحلیل امار توصیفی و مصور سازی

ابتدا سعی می کنیم بفهمیم چطور ویژگی ها هر ملک بر روی قیمت گذاری آن تاثیر خواهد گذاشت.



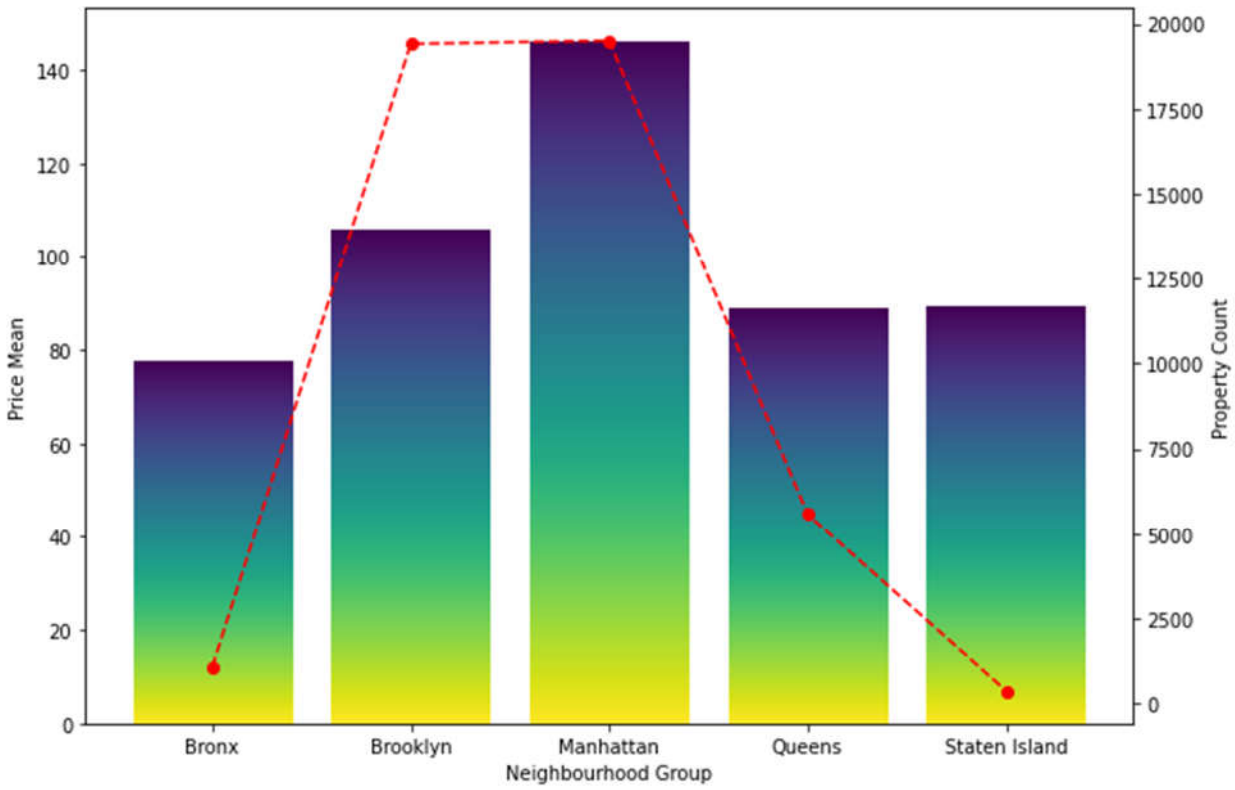
بر اساس نمودار بالا، می توان نتیجه گرفت که خانه های به اشتراک گذاشته شده در منطقه منهتن به طور معمول قیمت بیشتری نسبت به باقی مناطق دارد.

حال جدولی از امار توصیفی مرتبط با قیمت بر اساس مناطق مختلف را ایجاد می کنیم.

همانطور که درجدول زیر مشاهده می کنیم، میانگین قیمت در منطقه منهتن از باقی مناطق بیشتر است و همچنین پراکندگی قیمت ها هم بیشتر است چرا که انحراف معیار بیشترین مقدار را دارد.

	price							
	count	mean	std	min	25%	50%	75%	max
neighbourhood_group								
Bronx	1069.0	77.4	47.1	10.0	45.0	65.0	95.0	325.0
Brooklyn	19406.0	105.7	60.9	10.0	60.0	90.0	140.0	333.0
Manhattan	19500.0	145.9	70.4	10.0	90.0	135.0	199.0	333.0
Queens	5567.0	88.9	53.5	10.0	50.0	74.0	108.0	325.0
Staten Island	365.0	89.2	57.7	13.0	50.0	75.0	105.0	300.0

حال نمودار میله ای برای میانگین قیمت ها و تعداد ملک های به اشتراک گذاشته رسم می کنیم تا درک بهتری از این موضوع داشته باشیم.



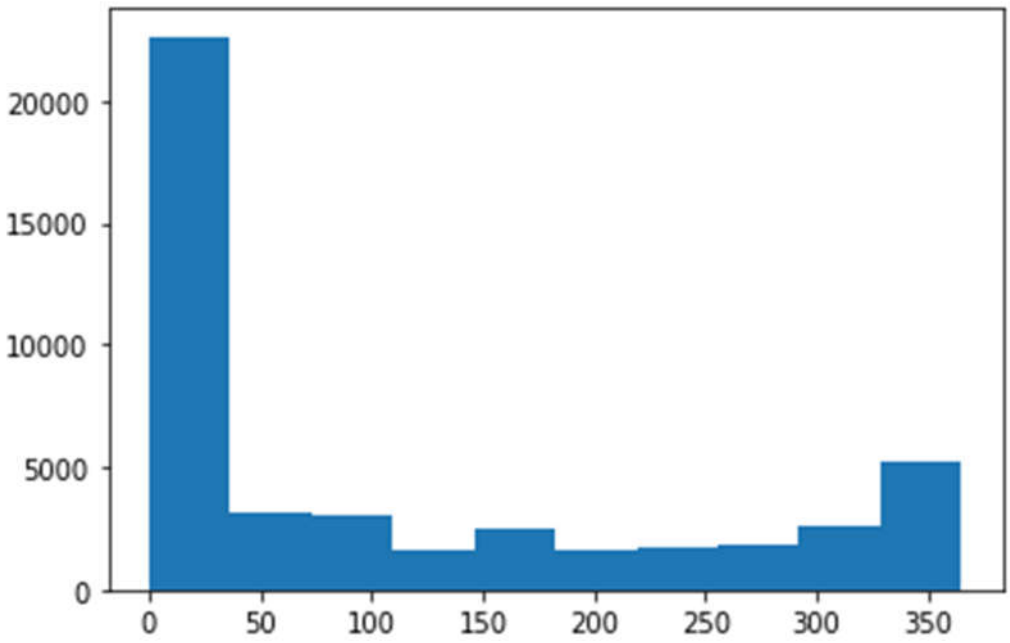
همانطور که در نمودار پراکندگی دیدیم، منهتن قیمت های بالاتری دارد و میانگین برای این منطقه این را نشان می دهد. با این حال، 4 منطقه دیگر نزدیکتر هستند. بروکلین دومین منطقه گران قیمت است. 3 منطقه دیگر تفاوتی ندارد.

از آنجا که ما قیمت های بالا را با IQR کاهش دادیم، حداکثر قیمت ها در همه مناطق نزدیک است. اجازه دهید به آمار توصیفی در داده های خام نگاه کنیم.

	price							
	count	mean	std	min	25%	50%	75%	max
neighbourhood_group								
Bronx	1091.0	87.5	106.7	0.0	45.0	65.0	99.0	2500.0
Brooklyn	20104.0	124.4	186.9	0.0	60.0	90.0	150.0	10000.0
Manhattan	21661.0	196.9	291.4	0.0	95.0	150.0	220.0	10000.0
Queens	5666.0	99.5	167.1	10.0	50.0	75.0	110.0	10000.0
Staten Island	373.0	114.8	277.6	13.0	50.0	75.0	110.0	5000.0

همانطور که می بینیم، تفاوت زیادی در Staten Island قبل و بعد از رسیدگی به موارد پرت داریم. تفاوت بین تعداد داده ها ها حدود 8 است، اما تفاوت بین میانگین، std و حداکثر قیمت بسیار زیاد است. این نشان می دهد که ما خروجی های واقعی را حذف کرده ایم. با این حال، ما مقدار زیادی از داده ها را در منهتن از دست دادیم. این موضوع منطقی است، زیرا ما در روش IQR خود فقط از کران بالا استفاده کردیم و منهتن گران‌ترین قسمت شهر است. بنابراین، ما رکوردهای بیشتری را از منهتن حذف کردیم.

حال نگاهی به توزیع abvailibity_365 خواهیم داشت.



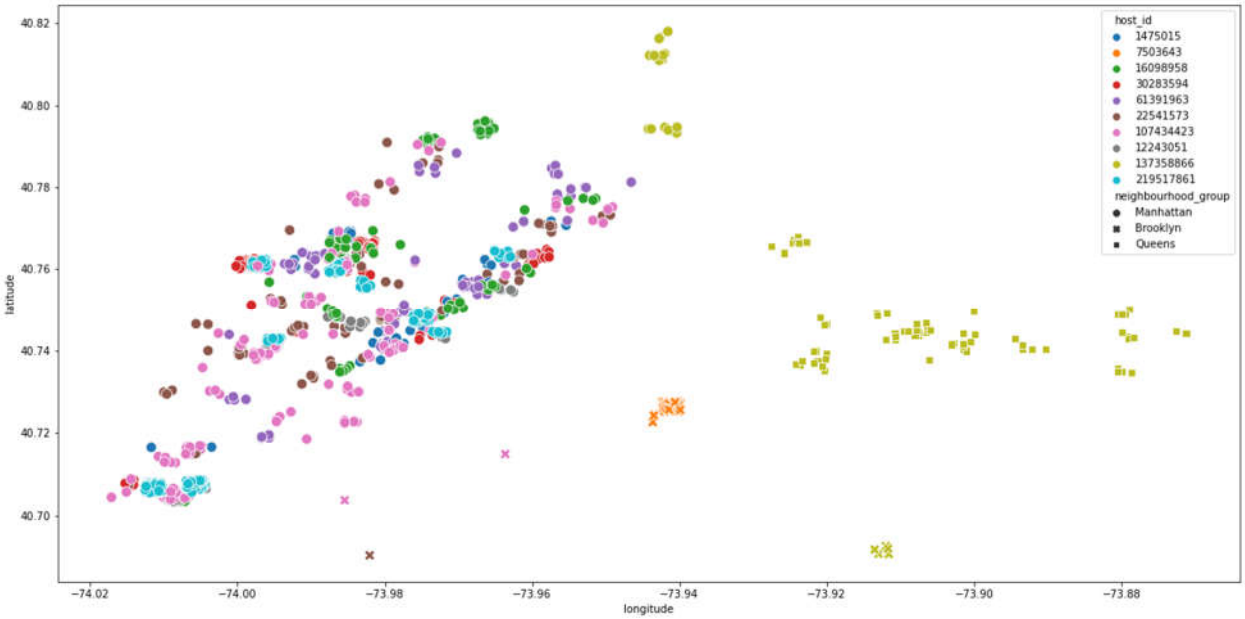
خانه هایی که کمتر از 36 روز در دسترس هستند شامل اکثر موارد می شوند. سایر بازه ها احتمال رخداد یکسانی را دارند.

3-2- آنالیز میزبانان

در این بخش سعی می کنیم پیدا کنیم که کدام مالک بیشترین ملک ها ر به اشتراک گذاشته و ببینیم آیا می توانیم هر نوع الگوی را در این مورد پیدا کنیم.

```
219517861    272
107434423    176
137358866    103
30283594      95
12243051      95
             ...
69254072       1
20354912       1
10777637       1
69236458       1
68119814       1
Name: host_id, Length: 35386, dtype: int64
```

براساس نتایج بالا 35386 میزبان منحصر به فرد وجود دارد. حال به 10 میزبان با بیشترین ملک به اشتراک گذاشته خواهیم داشت.



نمودار بالا نشان میدهد که این 10 میزبان در مناطق منهتن، بروکلین و کوییز فعالیت می‌کنند.

با توجه به نمودار فوق می توان نتیجه گرفت که میزبان شماره 137358866 در قسمت شرقی شهر کار می کند. علاوه بر این، او میزبان اصلی در منطقه کوئینز است. میزبان شماره 7503643 فقط روی بروکلین کار می کند و او در این منطقه مسلط است. دیگر میزبانان برتر، روی منهتن کار می کنند.

حال به تعداد کامنت‌ها در هر ماه برای این 10 میزبان نگاهی خواهیم داشت.

	reviews_per_month							
	count	mean	std	min	25%	50%	75%	max
host_id								
107434423	176.0	0.030	0.080	0.0	0.000	0.000	0.000	0.39
12243051	95.0	0.091	0.181	0.0	0.000	0.000	0.090	0.79
137358866	103.0	0.220	0.283	0.0	0.000	0.000	0.380	1.00
1475015	52.0	0.071	0.061	0.0	0.030	0.050	0.110	0.23
16098958	90.0	0.079	0.119	0.0	0.000	0.045	0.108	0.77
219517861	272.0	1.089	1.140	0.0	0.000	0.865	1.935	4.52
22541573	87.0	0.049	0.102	0.0	0.000	0.000	0.060	0.73
30283594	95.0	0.034	0.107	0.0	0.000	0.000	0.050	1.00
61391963	91.0	0.233	0.181	0.0	0.125	0.220	0.300	1.00
7503643	52.0	0.095	0.056	0.0	0.060	0.085	0.120	0.28

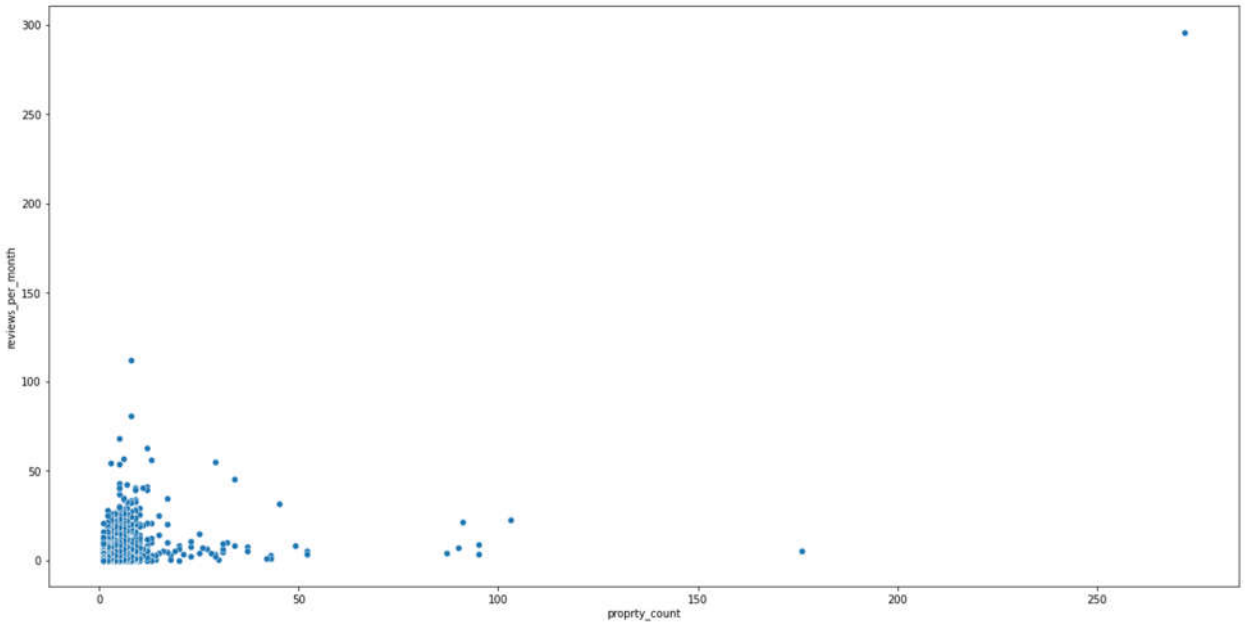
همانطور که در جدول بالا می بینیم میزبان شماره 219517861 که دارای بزرگترین سبد از املاک است، همچنین دارای بیشترین بررسی در هر ماه است. به نظر میرسد، برای هیچ یک از این میزبان ها، مدیریت املاک تنها توسط یک نفر انجام نمی شود. بنابراین، من فکر می کنم شرکت های کوچکی وجود دارند که املاک خود را در سایت Airnbn مدیریت می کنند. با این حال، برای شماره میزبان 219517861 همه چیز کمی متفاوت است. از نظر من، آنها تیم بزرگی دارند، با تیم بازاریابی عالی. احتمالا آنها ارتباط خوبی با مهمانان خود دارند و آنها را متقاعد کنند که نظرات خود را ارسال کنند، یا اگر بخواهیم کمی بدبین باشیم، ممکن است کامنت ها درباره خانه‌های به اشتراک گذاشته توسط کارکنان خود آنها ارسال شده تا آنها را در لیست برترین خانه ها بر اساس تعداد بررسی قرار دهند.

4-2- بررسی تاثیر تعداد کامنت ها برای هر ملک و عوامل آن

ابتدا یک دیتا فریم جدید ایجاد می کنیم تا تجمیع اطلاعات مرتبط با تعدادکامنت ها، تعداد املاک هر میزبان، قیمت و میانگین قیمت را ذخیره کند.

	host_id	proprty_count	price	reviews_per_month	avg_price
0	219517861	272	56166	296.11	206.5
1	107434423	176	49149	5.35	279.3
2	137358866	103	4514	22.68	43.8
3	30283594	95	20885	3.26	219.8
4	12243051	95	20074	8.65	211.3

حال نمودار پراکندگی مرتبط با تعداد کامنت ها و تعداد املاک را رسم می کنیم.



از نمودار بالا تحلیل خاصی را نمی توان انجام داد.

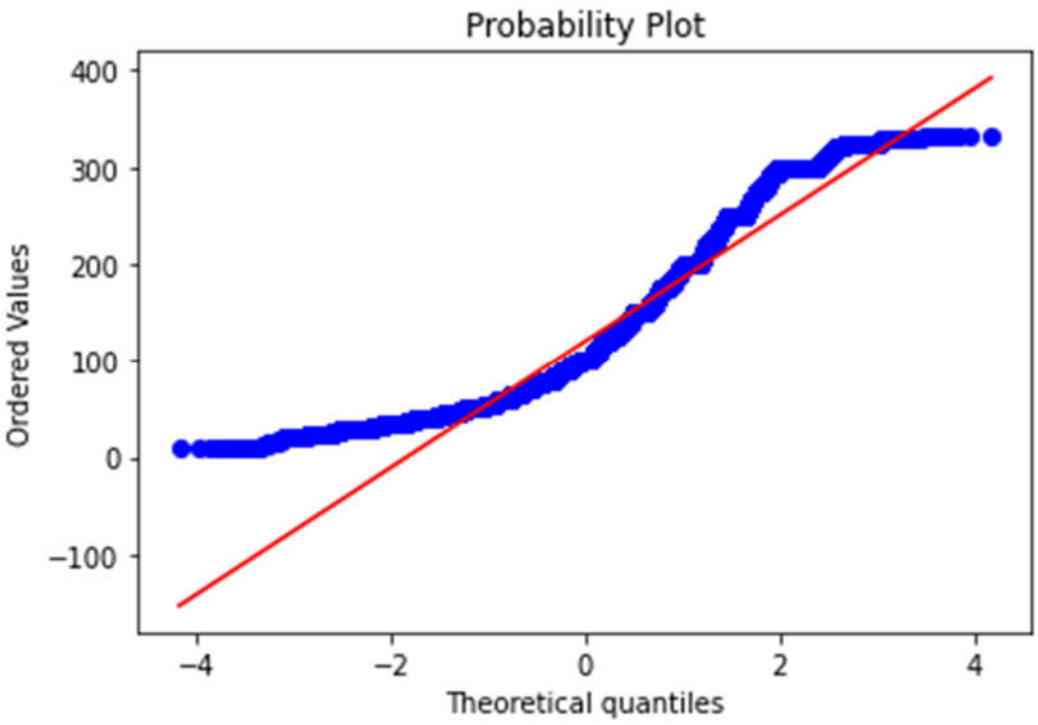
5-2- تعریف 5 آزمون فرض

در این بخش 5 آزمون فرض متفاوت رو بررسی می کنیم.

H1 -2-5-2: آیا توزیع داده های قیمت توزیع نرمال دارند؟

برای بررسی نرمال بودن داده‌ها از متد normaltest در کتابخانه Scipy استفاده می کنیم. مقدار p-value برای این آزمون فرض مقدار صفر است. درنتیجه فرض صفر مبنی نرمال بودن داده ها رد خواهد شد.

حال نمودار احتمال نرمال را هم رسم می کنیم تا بررسی دوباره ای داشته باشیم.



نمودار بالا همچنین نشان می دهد که داده های قیمت از یک خط مستقیم پیروی نمی کنند. بنابراین داده های قیمت توزیع نرمال ندارد. حال توزیع نرمال را در هر منطقه هم بررسی کنیم.

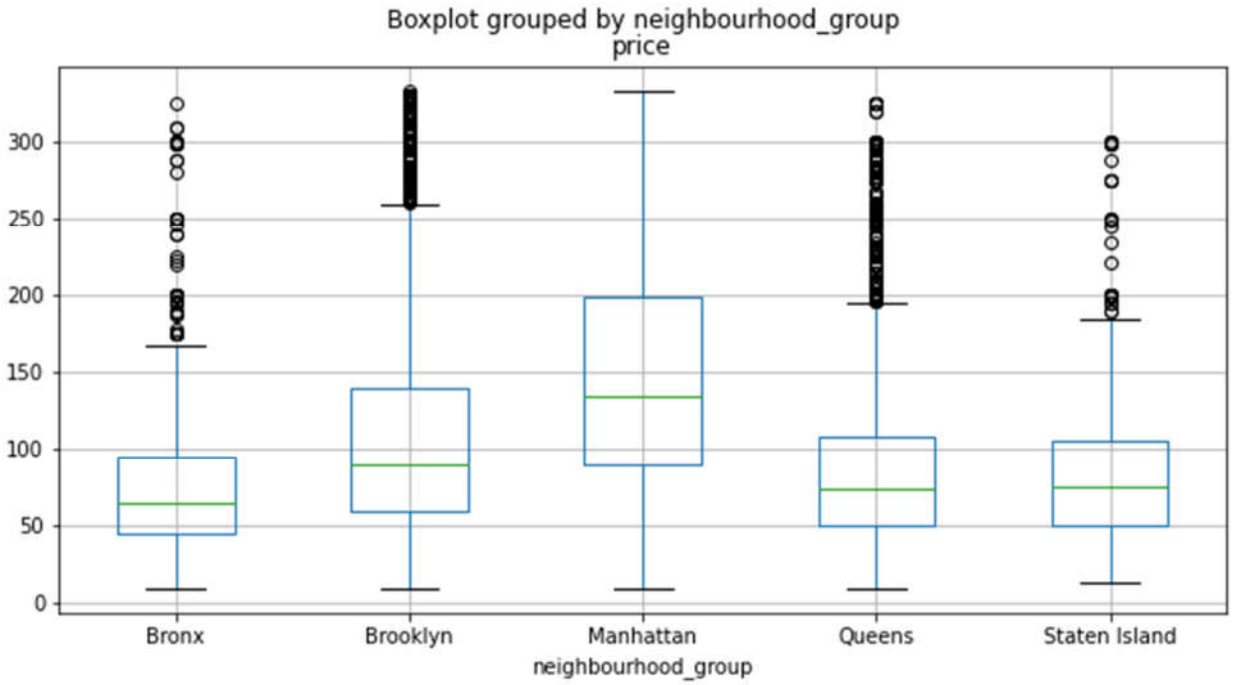
	neighbourhood_groop	statistic	p_value
0	Brooklyn	3293.48	0.0
1	Manhattan	1273.62	0.0
2	Queens	1689.98	0.0
3	Staten Island	125.64	0.0
4	Bronx	481.67	0.0

مقدار p-value برای تمامی این مناطق هم برابر با صفر است. پس قیمت در هیچ یک از این مناطق هم توزیع نرمال ندارد.

2-5-2 H2: آیا میانگین قیمت در هر منطقه با هم برابر است؟

برای بررسی این موضوع از آنالیز واریانس یک طرفه استفاده می‌کنیم.

ابتدا یک باکس پلات برای قیمت در هر منطقه رسم می کنیم.



بر اساس نمودار بالا تفاوت هایی در قیمت برای مناطق بروکلین، منهتن و سایر مناطق وجود دارد. بگذارید ببینیم که تست ANOVA این موضوع را ثابت می کند یا نه.

```
F_onewayResult(statistic=1506.07670402143, pvalue=0.0)
```

مقدار p-value برابر صفر است. در نتیجه قیمت در این مناطق با هم برابر نیست و فرض صفر مبنی بر برابر بودن میانگین قیمت در این مناطق رد می شود(البته باید توجه داشت که آنالیز واریانس زمانی قابل اجرا است که داده ها توزیع نرمال دارند و ما به این نتیجه رسیدیم که داده های قیمت از توزیع نرمال پیروی نمی کنند. با این حال ما این فرض را ریلکس کردیم تا فقط تستی بر اساس ANOVA داشته باشیم.

3-5-2: H3: آیا میانگین قیمت در دو منطقه Bronx و Staten Island برابر هستند؟

ابتدا باید ببینیم داد ها واریانس برابر دارند یا نه. از آنجایی که نتیجه گرفتیم قیمت در مناطق مختلف توزیع نرمال ندارد، بنابراین بهتر است از آزمون لوون برای برابری واریانس ها استفاده کنیم. برای توضیحات بیشتر می توان به اینجا مراجعه کرد.

```
LeveneResult(statistic=10.8702277774617, pvalue=0.0010011708090621695)
```

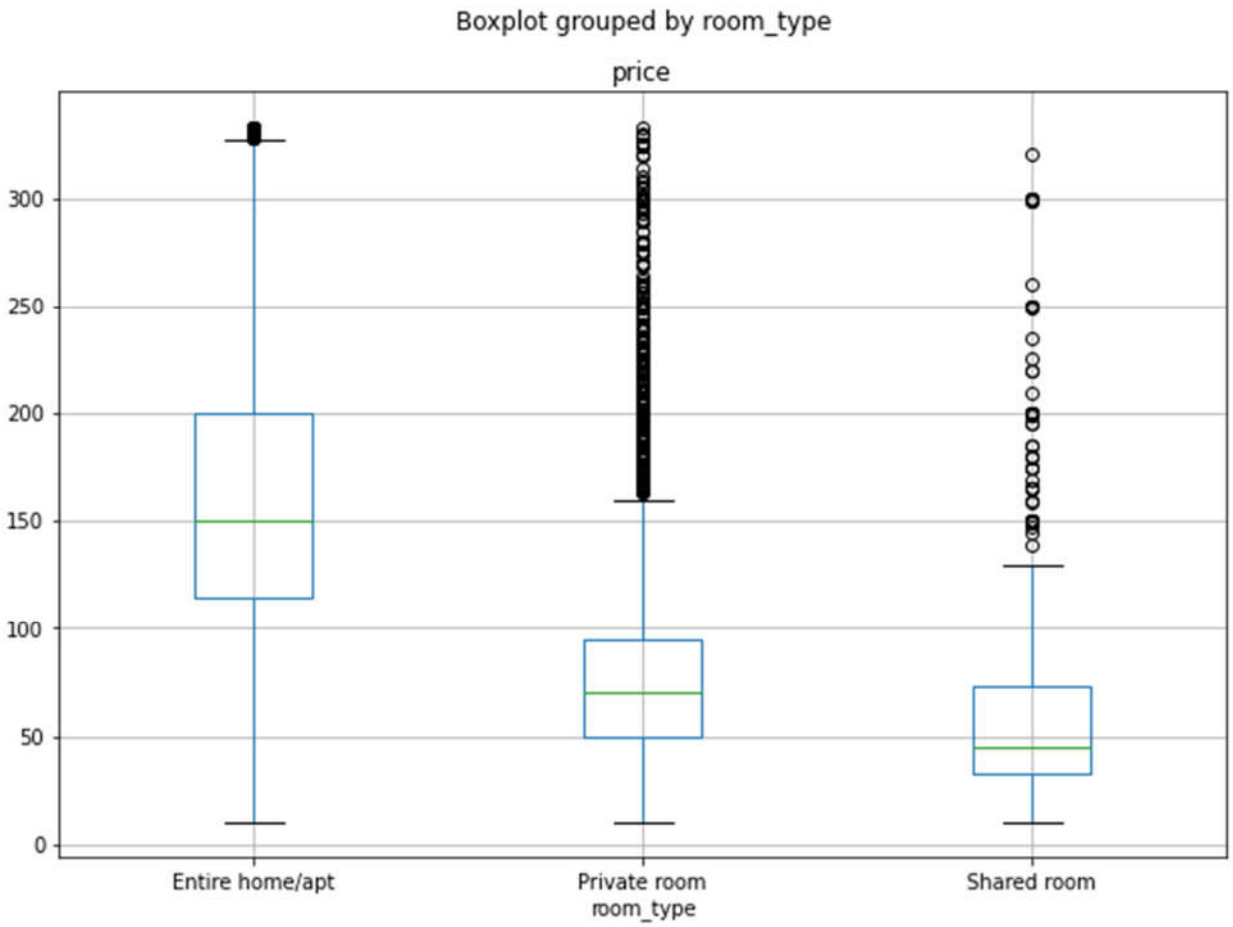
مقدار P-value برای این آزمون نشان میدهد که واریانس ها برابر نیستند. حال با استفاده از آزمون t در شرایط نابرابری واریانس ها، فرض برابری میانگین ها در این دو منطقه را برررسی می کنیم.

```
Ttest_indResult(statistic=-3.5261764309472103, pvalue=0.00045753711043566527)
```

مقدار p-value برای این دومقدار هم نشان میدهد که میانگین در این دور منطقه برابر نیستند.

H4 -2-5-4: میانگین قیمت خانه ها در private-room ها بیشتر از shared-room هاست؟

ابتدا یک باکس پلات برای مقایسه قیمت بر اساس نوع اتاق ها رسم می کنیم.



نمودار بالا می دهد که تفاوت های جزئی در قیمت بین اتاق خصوصی و اتاق مشترک وجود دارد. اما باید مطمئن باشیم که این تفاوت معنادار است.

با این حال، آزمون t در کتابخانه scipy دو طرفه است است. برای حل ین محدودیت، آزمون مقدار آماره t را محاسبه می کنیم و سپس آن را با مقدار t مورد انتظار مقایسه می کنیم تا فرضیه صفر را بپذیریم.

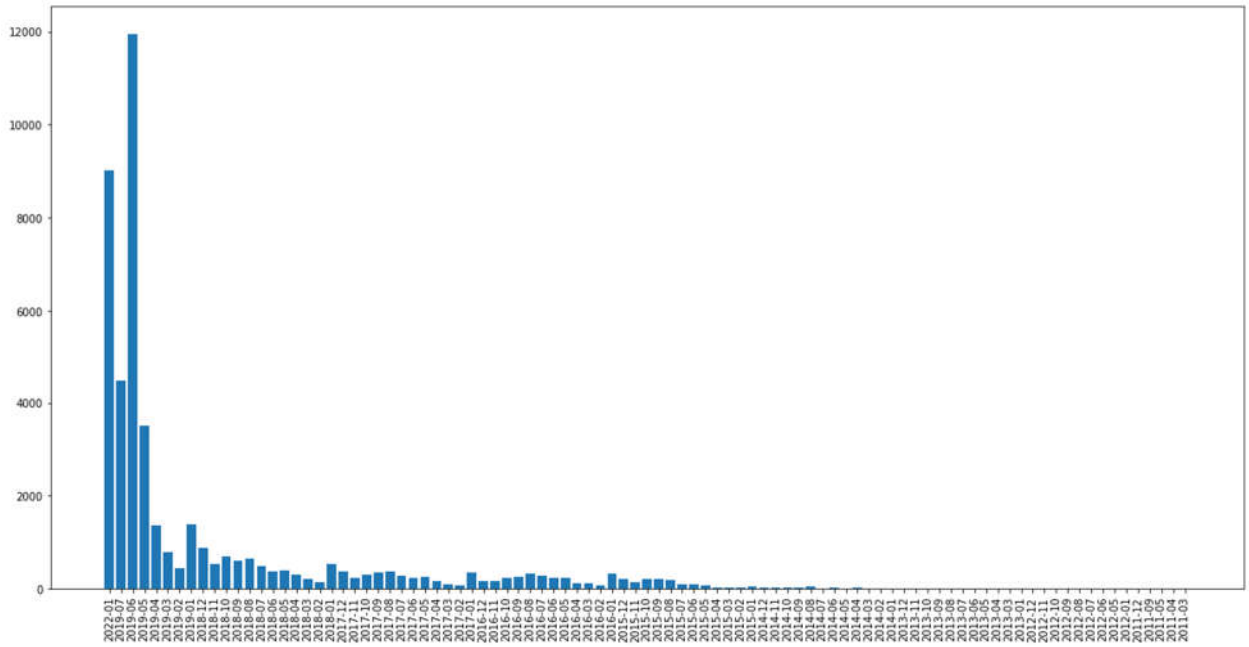
```
(t_crit, t_value)
```

```
(1.6449195280269775, -15.656058001740146)
```

t_crict بیشتر از t_value است، بنابراین، می‌توانیم فرضیه صفر را بپذیریم.

5-5-2- H5: آیا تعدا اخرین کامنت ها در هر ماه توزیع نمایی دارد؟

ابتدا نمودار میله ای تعداد اخرین کامنت ها بر اساس ماه را رسم می‌کنیم.



با توجه به نمودار بالا مشاهده می کنیم که تعداد آخرین بررسی برای هر ماه به صورت تصاعدی رشد می کند. با این حال، برای دو ماه آخر کاهش یافته است. برای سال 2022-01، ما داده های از دست رفته را به صورت دستی درج کردیم، بنابراین باید این رکوردها را حذف کنیم. برای 2019-07، به نظر این اتفاق افتاده که زمان استخراج داده ها شاید در اواسط ماه 2019-01 بود، بنابراین این ماه همه داده ها را ندارد. بنابراین، ما این رکوردها را نیز حذف خواهیم کرد.

حال از تست کولوموگروف-اسمیرنوف برای نیکویی برازش داده های مبتنی بر توزیع نمایی استفاده می کنیم.

```
KstestResult(statistic=0.3496493313021043, pvalue=1.8919859471039295e-10)
```

مقدار p-value نشان می دهد که نمی توان فرض صفر را پذیرفت و فرض پیروی از توزیع نرمال رد می‌شود.

6-2- پیش بینی قیمت اتاق‌ها بر اساس فیچرهای موجود

برای آماده سازی داده ها به منظور ارائه به مدل های رگرسیونی، ابتدا ستون های که لازم نیست را حذف می کنیم. سپس داده های کتگوریکال را به صورت dummy variable ترنسفورم می نماییم.

سپس داده ها را به دو دسته تست و ترین تقسیم خواهیم کرد.

ما برای اینکه اثر نرمال کردن داده های عددی را ببینیم، مدل را دو حالت داده های نرمال شده و نشده آموزش میدهیم.

علاوه بر این از دو روش رگرسیون خطی و Random Forest Regressor استفاده می کنیم.

مقادیر R2 برای این روشها در جدول زیر آمده است:

	داده های اصلی	داده های نرمال شده
Linear Regression	0.1056	0.1228
Random Forest Regressor	**0.5673**	**0.3212**

براساس جدول بالا رگرسیون خطی بر روی داده های نرمال شده بهتر عمل می کند. حال آنکه random fores بر روی داده های اصلی نتیجه بهتری دارد.

2-7- تسک امتیازی یک

در این مرحله می خواهیم ببینیم تاثیر نزدیکی امکانات در اطراف یک اتاق به اشتراک گذاشته شده در قیمت ان چگونه است؟

2-7-1 قدم اول: خواندن دیتا مرتبط با مکان امکانات در شهر نیویورک

ما اطلاعات مرتبط با موقعیت مکانی ایستگاه های مترو و همچنین، فرودگاه کندی را به مدل اضافه می کنیم.

به علت تکراری بودن بعضی از ایستگاه های مترو، مقادیر تکرار شده را حذف می‌کنیم.

2-7-2 قدم دوم: محاسبه فاصله هر خانه تا ایستگاه های مترو و فرودگاه

نحوه محاسبه این فاصله از اینجا تطبیق داده شده است. علاوه بر این به علت محدودیت زمانی، تابع توسعه داده شده در این لینک استفاده شده است.

بعد از این مرحله، برای هر خانه موقعیت نزدیک ترین مترو به آن پیداد شده ومقدار فاصله تا آن ایستگاه مترو در ستون جدیدی ثبت میشود.

2-7-3- بررسی رابطه بین قیمت و فاصله تا مراکز حمل و نقل

برای بررسی این مورد اینبار رگرسیون را با این فیچرهای جدید اجرا خواهد شد.

مقدار R2 با این قیچر جدید برابر با 0.58 شده که تغییرات چندانی نداشته. بنابراین می توان نتیجه گرفت که این اطلاعات جدید در بهبود مدل تاثیری نداشته اند.

2-8 بررسی جنسیت میزبان در قیمت گذاری و تعداد کامنت ها

در این مرحله می‌خواهیم بررسی کنیم که آیا رابطه‌ای بین جنسیت میزبان و سایر ویژگی‌های عددی مانند قیمت و نظرات وجود دارد یا خیر.

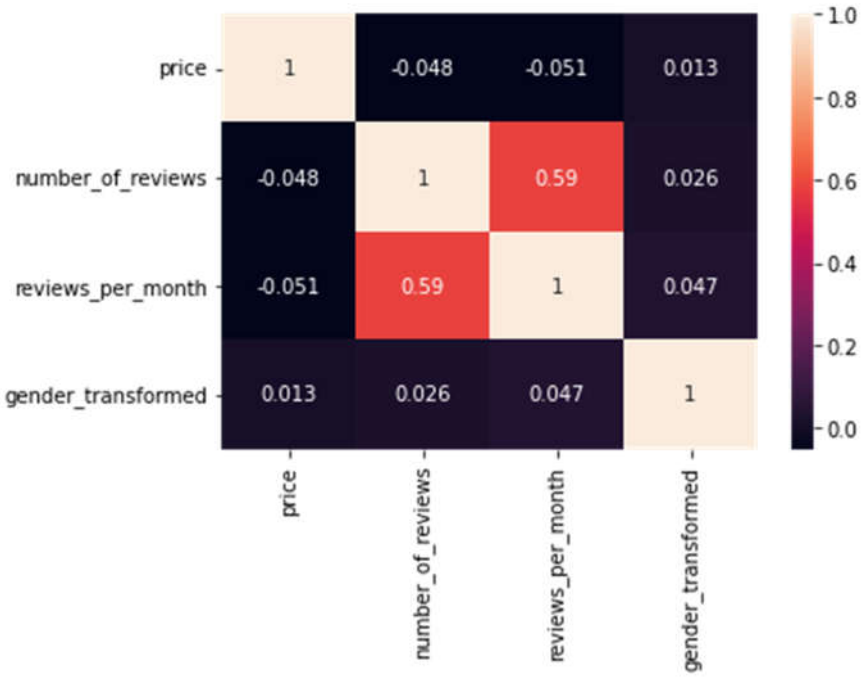
2-8-1 نصب و ایمپورت کردن کتابخانه های لازم

ابتدا کتابخانه gender_guesser برای پیش بینی جنسیت بر اساس نام میزبان نصب و ایمپورت می کنیم.

2-8-2- ترنسفورم کردن داده های مرتبط با جنسیت

در این مرحله می‌خواهیم بررسی کنیم که آیا رابطه‌ای بین جنسیت میزبان و قیمت، تعداد بررسی‌ها، بازبینی‌ها در ماه وجود دارد یا خیر.

برای بررسی آن، از ماتریس همبستگی استفاده می کنیم. اما قبل از آن، باید جنسیت را به مقادیر عددی تبدیل کنیم. در این مورد از onhotencoder استفاده می کنیم. همانطور که در ماتریس همبستگی پایین می بینیم هیچ ارتباطی بین جنسیت میزبان و سایر پارامترها وجود ندارد.



3- دیتا ست 2

این تمرین از دیتاست از آ گهی های استخراج شده از یکی از بزرگترین پلتفرم های املاک کشور آلمان استفاده میشود.

نسک های مرتبط با این پروژه به صورت زیر خواهد بود:

1- پاکسازی داده از قبیل بررسی داده های از غیرموجود و یافتن داده های پرت . بررسی موارد تکراری
2- رایه اطلاعات تجمیعی از تعداد آ گهی و پارامترهای مختلف آ گهی از قبیل تعداد آگهی ها در مناطق مختلف جغرافیایی ، اطلاعات در خصوص تعداد انواع خانه ها ، بررسی قیمت در مناطق مختلف جغرافیایی و ... . لازم است حتما در این بخش از بحث مصورسازی داده ها استفاده کنید و این اشکال ایجاد شده را تفسیر کنید.
3- تلاش جهت مدل سازی قیمت ها بر اساس پارامتر های مختلف آ گهی.
4- استفاده از بحثmultiprocessing در بخش پاکسازی و پیش پردازش داده ها و و بررسیruntime فرایندها.
5- استفاده ازdask وpyspark در بخش پاکسازی و پیش پردازش داده ها و بررسیruntime فرایندها.

تسک های امتیازی

1- استفاده از مهندسی ویژگی در جهت بهبود مدل ها.
2- استفاده ازdask وpyspark در بخش مدلسازی داده ها و مقایسه با حالت عدم استفاده از آنها

3-1- پاکسازی داده ها

بعد از خواندن داده ها از فایل csv، این داده ها در دیتافریمی به اسم raw_data ذخیره شده است. در ابتدا یک نگاه اجمالی به داده های موجود خواهیم داشت.

regio1	serviceCharge	heatingType	telekomTvOffer	telekomHybridUploadSpeed	newlyConst	balcony	picturecount	pricetrend	telekomUploadSpeed
Nordrhein_Westfalen	245.00	central_heating	ONE_YEAR_FREE	NaN	False	False	6	4.62	10.0
Rheinland_Pfalz	134.00	self_contained_central_heating	ONE_YEAR_FREE	NaN	False	True	8	3.47	10.0
Sachsen	255.00	floor_heating	ONE_YEAR_FREE	10.0	True	True	8	2.72	2.4
Sachsen	58.15	district_heating	ONE_YEAR_FREE	NaN	False	True	9	1.53	40.0
Bremen	138.00	self_contained_central_heating	NaN	NaN	False	True	19	2.46	NaN

سپس داده های تکراری با استفاده از دستور dtrop_duplicates حذف می کنیم.

در مرحله بعد نگاه می کنیم که در هر فیچر چند درصد داده ها گمشده اند چرا که حجم داده های گمشده بسیار زیاد است و برای درک بهتر مقادیر از درصد استفاده میکنیم.

regio1	0.00
serviceCharge	2.57
heatingType	16.68
telekomTvOffer	12.13
telekomHybridUploadSpeed	83.25
newlyConst	0.00
balcony	0.00
picturecount	0.00
pricetrend	0.68
telekomUploadSpeed	12.41
totalRent	15.07
yearConstructed	21.22
scoutId	0.00
noParkSpaces	65.39
firingTypes	21.19
hasKitchen	0.00
geo_bln	0.00
cellar	0.00
yearConstructedRange	21.22
baseRent	0.00
houseNumber	26.42
livingSpace	0.00
geo_krs	0.00
condition	25.47
interiorQual	41.91
petsAllowed	42.62
street	0.00
streetPlain	26.41
lift	0.00
baseRentRange	0.00
typeOfFlat	13.62
geo_plz	0.00
noRooms	0.00
thermalChar	39.62

به علت اینکه تعداد داده گمشده قابل توجه است، ستون هایی که بیشتر از 15 درصد داده گمشده دارند را حذف می کنیم.

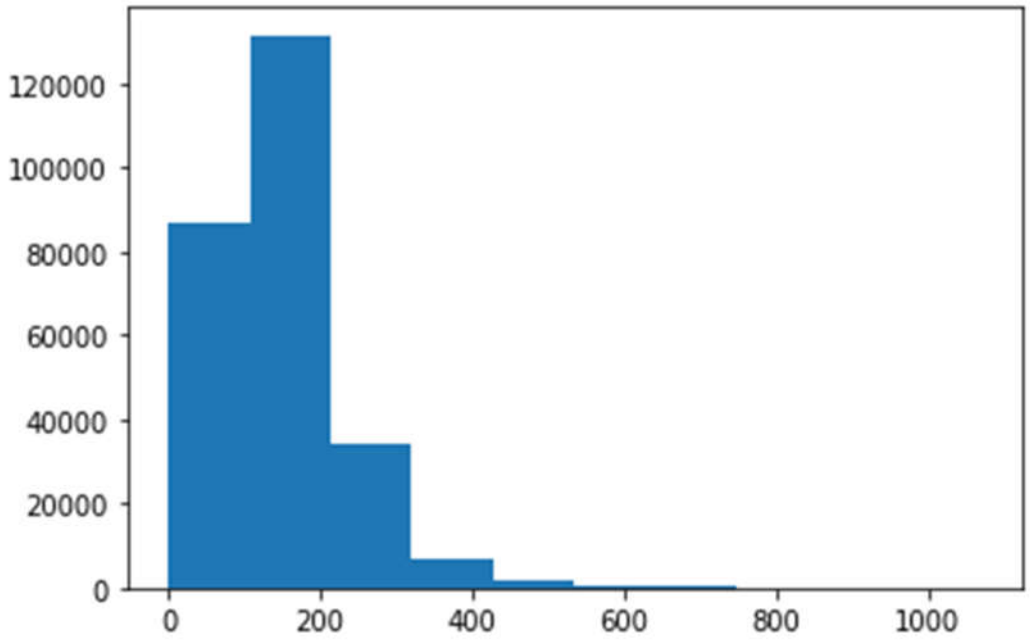
قبل از اصلاح داده های گمشده داده های پرت در هر فیچر را حذف می کنیم.

ابتدا با متد describe نگاهی به ستون serviceCharge خواهیم داشت.

```
count    261941.000000
mean        151.206113
std         308.295790
min           0.000000
25%          95.000000
50%         135.000000
75%         190.000000
max      146118.000000
Name: serviceCharge, dtype: float64
```

پراکندگی داده ها قابل توجه است. حال به استفاده از متد 3 سیگما داده های بزرگ تر از 3 سیگما را حذف می کنیم. به این ترتیب داده های گمشده این بخش هم حذف خواهد شد. متد دیگری که می توان استفاده کرد، جایگزیتی میانگین قیمت با داده های گمشده است.

نمودار هیستوگرام این فیچر بعد از حذف این داده به صورت زیر خواهد بود.



در ستون های typeOfFlat و telekomTvOffer مقدارهای گمشده را به طور کامل حذف می کنیم.

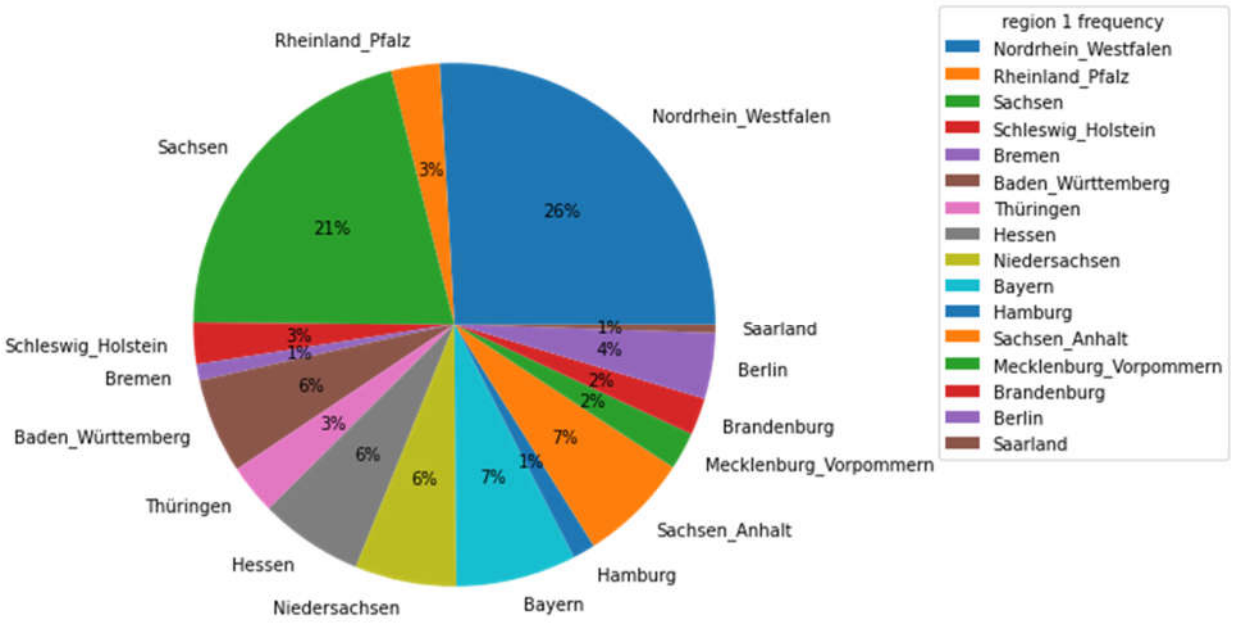
ستون description را هم به طور کامل حذف می کنیم.

در ستون های pricetrend و telekomUploadSpeed یا استفاده از متد interpolate داده های گمشده را ایمپیوت می کنیم.

به این ترتیب داده های گمشده در تمام دیتاست اصلاح می شوند و داده های پرت serviceCharge هم حذف می شوند. علاوه بر این می توان برای باقی مقادیر عددی هم از روشهایی مثل IQR یا 3σ استفاده کرد.

2-3- تفصیر داده ها

ابتدا درصد داده ها در هر regio1 را بررسی می کنیم. برای اینکار نمودار دایره ای بر اساس درصد تکرار هر منتطقه رسم می کنیم.

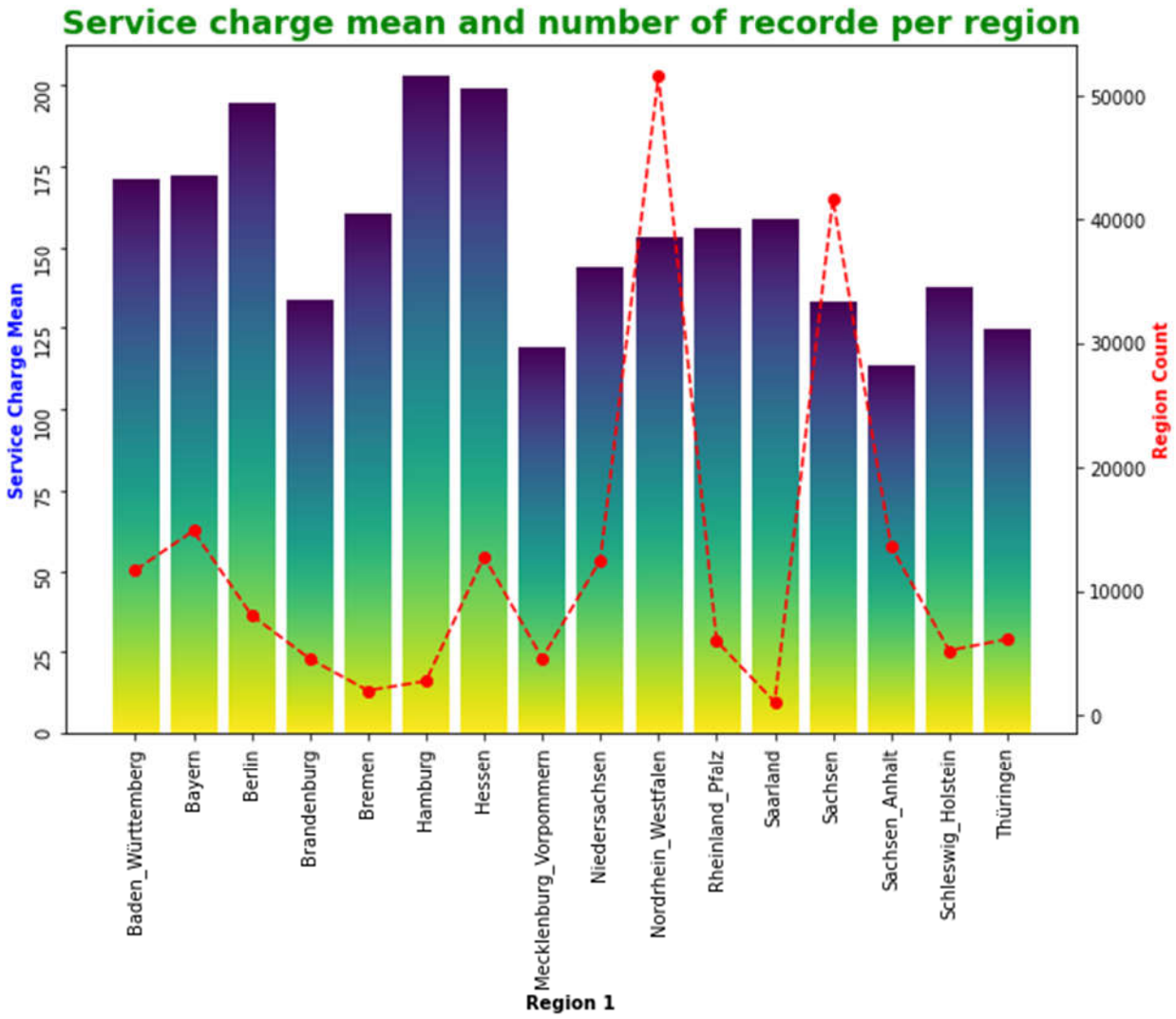


همانطور که مشاهده می شود دو منطقه نزدیک به 50 درصد حجم داده ها را به خود اختصاص داده اند.

حال به بررسی مقدار serviceCharge بر اساس مناطق مختلف می پردازیم.

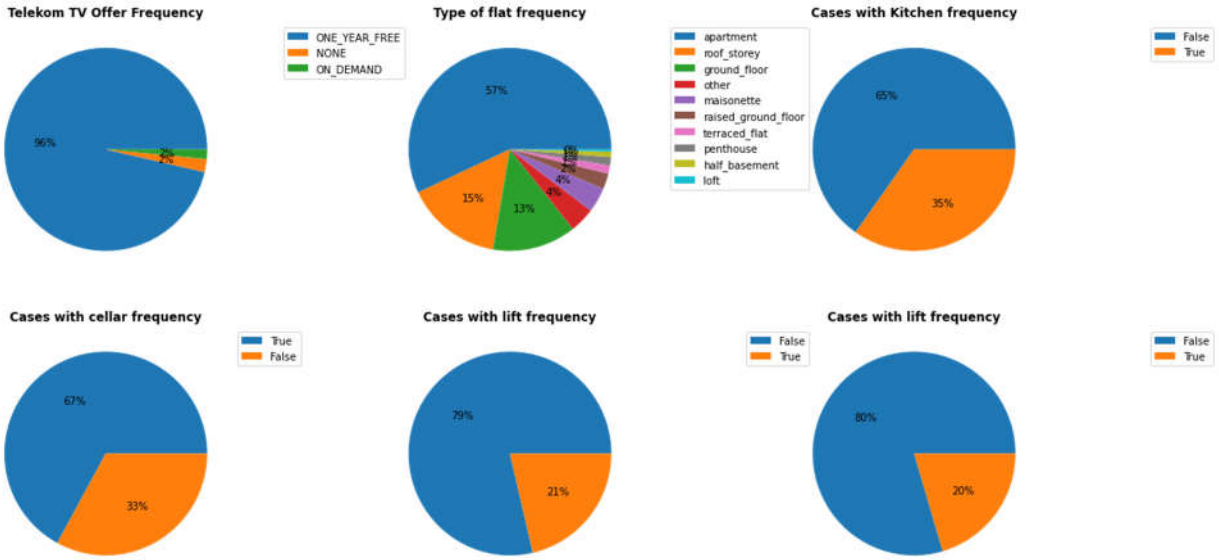
regio1	count	mean	std	min	25%	50%	75%	max
Baden_Württemberg	11683.0	170.42	90.91	0.0	110.00	160.0	220.00	830.00
Bayern	14870.0	171.52	89.62	0.0	110.00	160.0	210.00	980.00
Berlin	8081.0	194.09	125.25	0.0	111.00	164.0	236.00	1070.00
Brandenburg	4567.0	132.99	73.59	0.0	80.00	120.0	165.00	960.00
Bremen	2005.0	159.95	74.32	0.0	104.00	150.0	200.00	670.00
Hamburg	2763.0	202.36	123.98	0.0	115.00	175.0	250.00	1040.00
Hessen	12784.0	198.26	97.91	0.0	134.00	189.0	250.00	1000.00
Mecklenburg_Vorpommern	4606.0	118.80	61.50	0.0	73.61	110.0	150.00	603.13
Niedersachsen	12510.0	143.21	68.49	0.0	95.00	130.0	180.00	1000.00
Nordrhein_Westfalen	51616.0	152.82	76.88	0.0	100.00	137.0	185.00	1050.00
Rheinland_Pfalz	5989.0	155.25	70.39	0.0	100.00	150.0	200.00	950.00
Saarland	1080.0	158.42	69.72	0.0	100.00	150.0	200.00	500.00
Sachsen	41704.0	132.70	69.46	0.0	87.21	121.0	160.00	1009.56
Sachsen_Anhalt	13613.0	112.91	57.56	0.0	70.00	100.0	146.00	923.72
Schleswig_Holstein	5223.0	137.02	68.16	0.0	90.00	120.0	171.00	800.00
Thüringen	6142.0	124.39	61.51	0.0	80.00	120.0	150.17	992.25

از روی جدول بالا تحلیل کردن کمی مشکل است به خاطر تعدا بالای مناطق و تنوع در اعداد. به همین خاطر نمودار میله برای بررسی این موضوع رسم می کنیم.

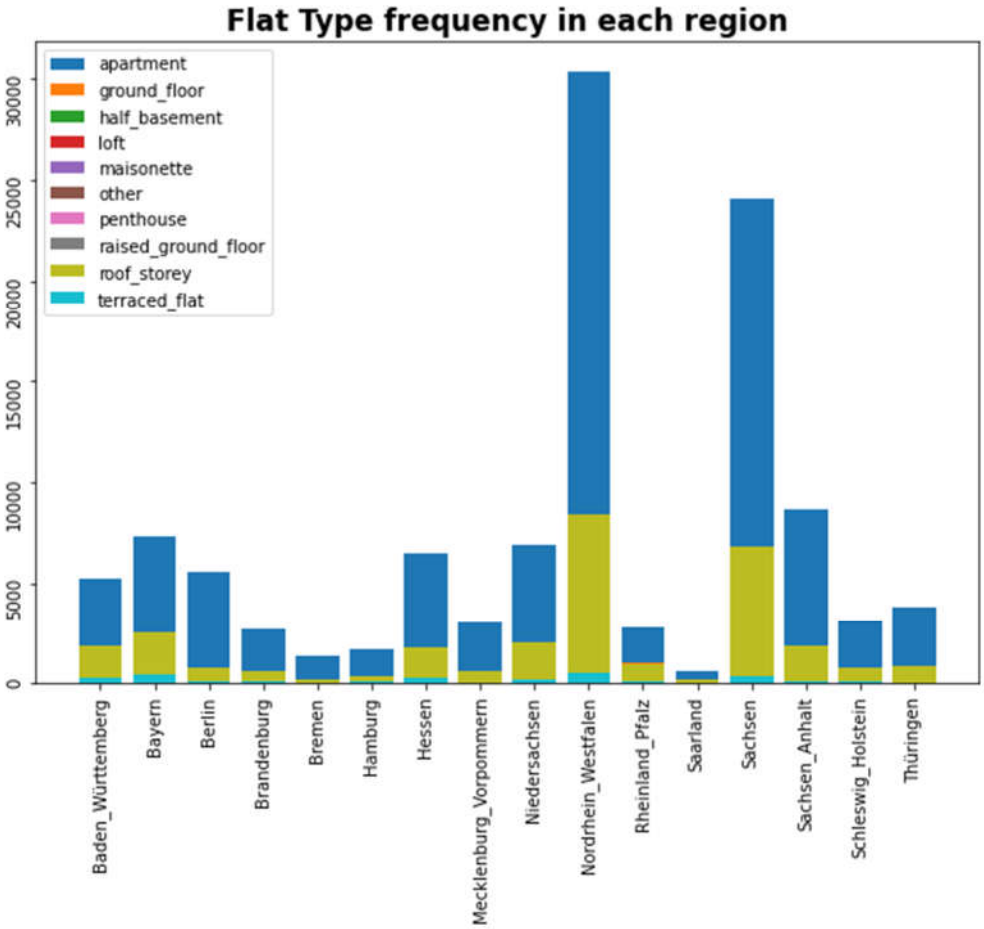


این نمودار نشان می دهد با اینکه موارد بیشتری در دو منطقه آلمان وجود دارد، ولی صرفا هزینه ها در این مناطق الگوی مشخصی بر اساس تعداد آگهی ندارد.

حال برای باقی ستون ها نمودار دایره ای رسم می کنیم.



همانطور که مشاهده می کنیم تنوع بالای در انواع آپارتمان ها(type_of_flat) وجود دارد. می خواهیم بررسی کنیم در مناطق مختلف آلمان، چه ترکیبی از این انواع فلت ها وجود دارد.



3-3- مدلسازی قیمت ها بر اساس پارامترهای آگهی

ابتدا داده ها را برای ورود با مدل یادگیری ماشین اماده سازی می کینم. برای سادگی کار داده ها را به 5 دسته boolean_feature، categorical_feature، float_feature و int_feature تقسیم می کنیم.

سپس داده های کتگوریکال و بولین را ترنسفورم می کنیم.

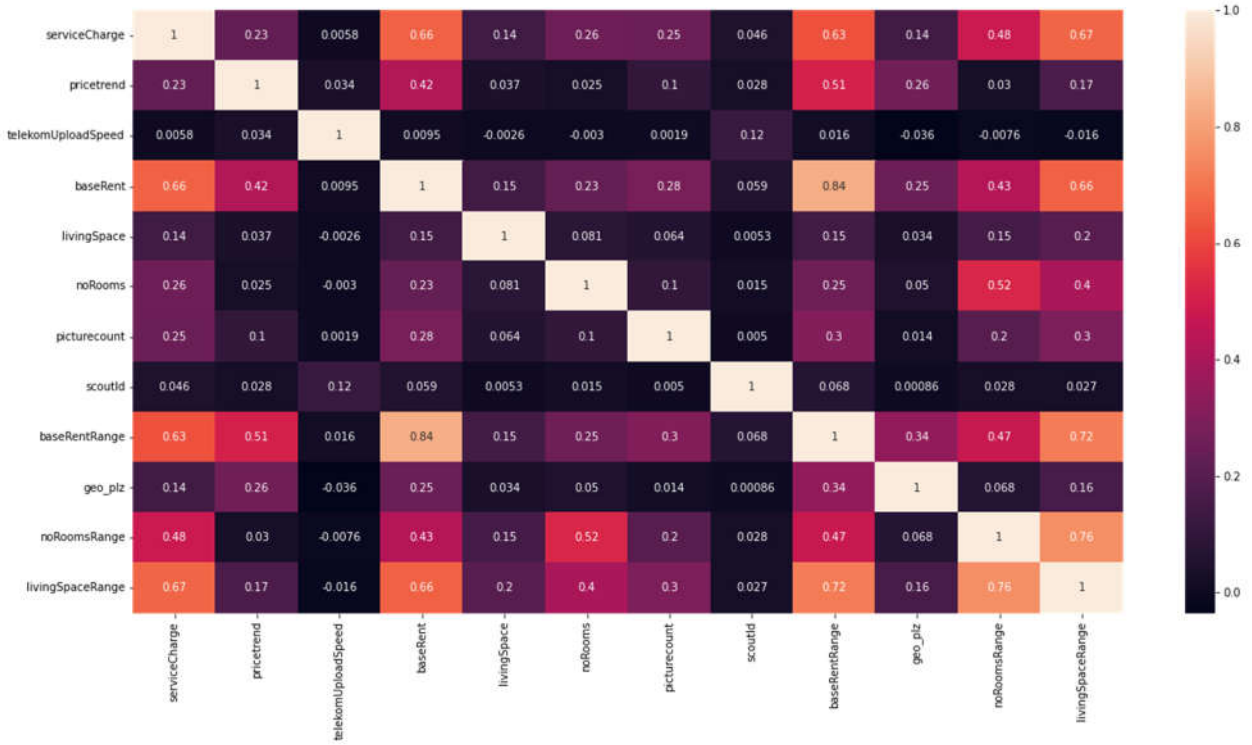
پس از آماده سازی بردارهای X و y آنها را به دست اموزشی و تست تقسیم کرده و مدل رگرسیونی مد نظر را ایمپورت می کنیم. حال مدل را ران می کنیم.

مشکلی بعد از ران کردن مدل پیش میاید این است که نوت بوک پیغام کرش را چاپ می کند. به نظر میرسد دلیل این اتفاق تعداد بیش از انداز ویژگی ها بعد از ترنسفورم کردن داده هاست.

به همین دلیل نیاز به مهندسی ویژگی ها خواهیم داشت تا فرایند بهتری را پیاده سازی کنیم.

3-5 تسک امتیازی 1: مهندسی ویژگی ها

به منظور بررسی روابط بین داده ها ابتدا ماتریس همبستگی بین داده های عددی را رسم می کنیم.



این نمودار نشان می دهد که همبستگی بالایی بین داده های عددی وجود ندار و بالاترین همبستگی 0.76 است که به نظر حذف ویژگی منطقی نباشد.

حال مدل رگرسیونی با استفاده از RandomForest را آموزش میدهیم.

مقدار MSE برابر با 2725.6 شده است. حال ویژگی هایی که همبستگی بالای 0.6 داشتند را حذف کرده ودوباره مدل را آموزش میدهیم تا ببینیم MSE تغییری می کند یا خیر.

این مقدار برابر 2848.9 شده است که تغییر چندانی نداشته است و نشان میدهد که فرایند مناسبی را طی نکردیم.