گزارش پروژه

علی صالح ۹۷۲۲۲۰۵۳

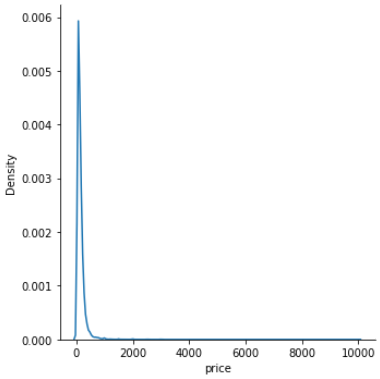
داده کاوی

پروژه سری اول

۱.۱- داده های new york airBNB

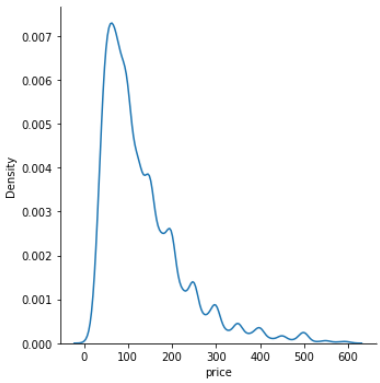
۱.۱.۱- پاکسازی داده ها:

- ابتدا ستون هایی که به آن نیاز نداریم را حذف می‌کنیم.
- در ادامه سعی می‌کنیم داده های outlier را حذف کنیم.
  برای این کار از فیچر price استفاده می‌کنیم.
  توزیع price:



داده هایی که قیمت آنها بیشتر از ۶۰۰ و کمتر
صفر است را حذف می‌کنیم.

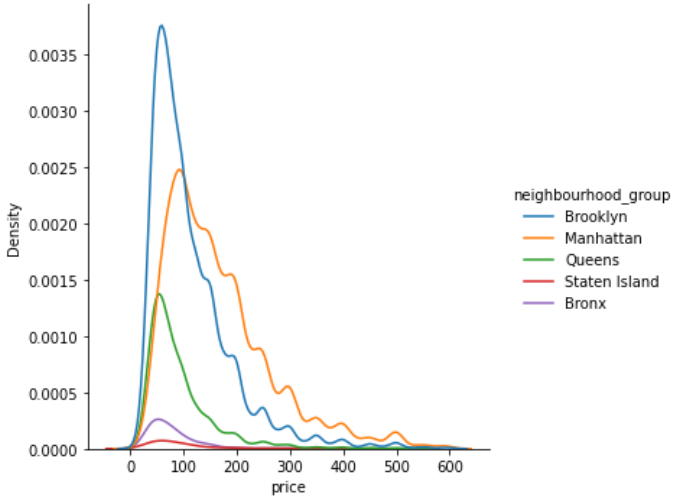
توزیع جدید price:



- بررسی داده های null
  این دیتاست دیتای null ندارد.

۱.۱.۲- ارائه اطلاعات کلی:

- قیمت خانه ها در نواحی مختلف را بررسی می‌کنیم.



neighbourhood_group	min	max	count
Bronx	10.0	500.0	1083
Brooklyn	10.0	599.0	19917
Manhattan	10.0	599.0	21019
Queens	10.0	545.0	5639
Staten Island	13.0	450.0	367

نرمال بودن توزیع قیمت در این نواحی را بررسی می‌کنیم

مقدار p-value این نواحی

0.0

5.018998498703473e-166

0.0

0.0

4.9426578184380965e-40

مشخص است که هیچکدام از این نواحی قیمت با توزیع نرمال ندارند.

مقدار log دیتا ها را بررسی می‌کنیم.

مقدار p-value:

4.2271434219973966e-68

4.6138862405883045e-15

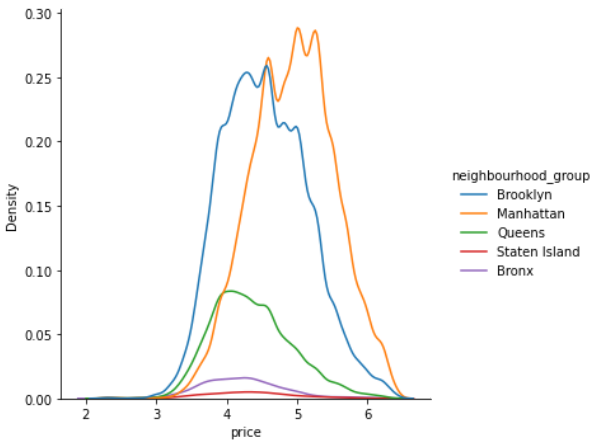
3.4851727210285935e-21

1.0245950120352195e-44

0.05015698771392723

با log transform داده ها به توزیع نرمال نزدیک ترند.

نمودار قیمت بعد از log transform:



بررسی عرضه تقاضا روی قیمت:

تعداد review ها را مقدار تقاضا در نظر می‌گیریم.

مجموع تعداد کامنت ها و تعداد خانه ها در جدول روبرو برای هر ناحیه مشخص است.

neighbourhood_group	count	sum
Bronx	1083	28313
Brooklyn	19917	484359
Manhattan	21019	448591
Queens	5639	156781
Staten Island	367	11540

نسبت تقاضا به عرضه را با تقسیم sum به count به دست می‌آوریم:

Bronx	0.038251
Brooklyn	0.041120
Manhattan	0.046856
Queens	0.035967
Staten Island	0.031802

همچنین میانگین قیمت در هر منطقه را هم بدست می‌آوریم:

Bronx	**4.227838**
Brooklyn	**4.545307**
Manhattan	**4.941573**
Queens	**4.358609**
Staten Island	**4.328403**

با حساب کردن کورلیشن متوجه میشویم نسبت تقاضا به عرضه رابطه مستقیم با میانگین قیمت دارد
همچنین از آنجایی که نسبت این دو تقریبا یکسان است (جدول پایین) متوجه می‌شویم این رابطه تقریبا خطی است.

Bronx	110.528881
Brooklyn	110.536737
Manhattan	105.463877
Queens	121.182308
Staten Island	136.102932

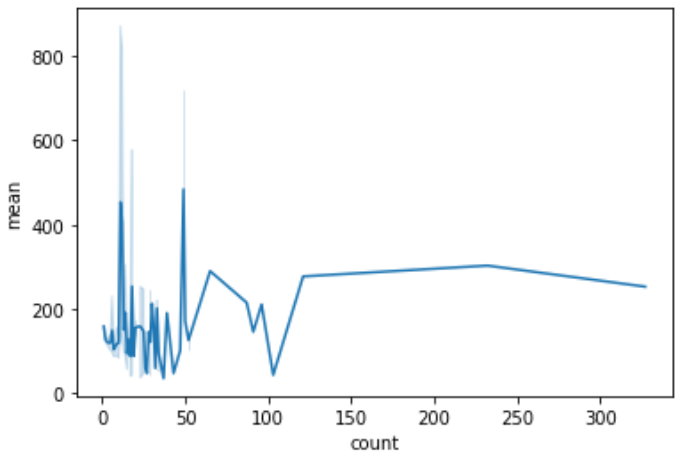
۱.۱.۳ و ۱.۱.۴ بررسی صاحبان آگهی

صاحبان آگهی را بررسی می‌کنیم.
دو عامل را برای اینکه بفهمیم چه کسانی بیشترین آگهی را دارند بررسی می‌کنیم

- قیمت

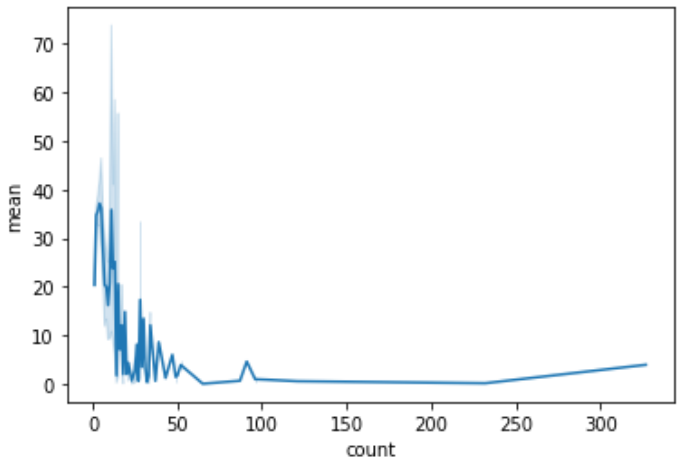
تعداد و میانگین قیمت آگهی هر صاحب خانه را در نظر می‌گیریم:

host_id	count	mean
2438	1	95.000000
48823036	1	180.000000
48819868	1	200.000000
48818023	1	78.000000
48817598	1	145.000000
...	...	...
16098958	96	208.958333
137358866	103	43.825243
30283594	121	277.528926
107434423	232	303.150862
219517861	327	253.195719



- تعداد کامنت ها

host_id	count	mean
2438	1	1.000000
48823036	1	156.000000
48819868	1	0.000000
48818023	1	0.000000
48817598	1	2.000000
...	...	...
16098958	96	1.437500
137358866	103	0.844660
30283594	121	0.537190
107434423	232	0.125000
219517861	327	3.917431



تعداد کامنت ها نسبت به قیمت کورلیشن بیشتری با تعداد آگهی های هر صاحب خانه دارد

۱.۱.۵- مطرح کردن آزمون فرض

- آیا قیمت خانه در منهتن از بروکلین گران تر است؟
فرض صفر: قیمت خانه در این دو منطقه یکسان است.
از t-test استفاده می‌کنیم.
مقدار p-value کمتر از alpha است پس فرض صفر رد می‌شود و نتیجه می‌گیریم قیمت خانه در این دو منطقه یکسان نیست.
و از آنجا که مقدار t-statistics بزرگتر از صفر است نتیجه می‌گیریم قیمت خانه در منهتن از بروکلین از آماری بیشتر است.
به همین شکل با مقایسه ی بقیه محله ها و طرح ۵ آزمون فرض دیگر در مورد منطقه های Queens Bronx Staten Island به نتایج زیر میرسیم:

Bronx < Staten Island = Queens < Brooklyn < Manhattan

فرض صفر فقط در t-test بین Staten Island و Queens رد نشد و در بقیه تست ها رد شد.

- آیا قیمت اتاق کامل از قیمت اتاق شخصی بیشتر است؟
فرض صفر: قیمت خانه در این دو تایپ یکسان است.
از t-test استفاده می‌کنیم.
مقدار p-value کمتر از alpha است پس فرض صفر رد می‌شود و نتیجه می‌گیریم قیمت خانه در این دو تایپ یکسان نیست.
و از آنجا که مقدار t-statistics بزرگتر از صفر است نتیجه می‌گیریم قیمت اتاق کامل از اتاق شخصی بیشتر است.

به همین شکل به این نتیجه می‌رسیم که قیمت اتاق شخصی از اتاق مشترک بیشتر است.

Shared room < private room < entire room

- آیا قیمت با **availability** همبستگی دارد؟
  از تست پیرسون استفاده می‌کنیم
  مقدار **p-value** از **alpha** کمتر است پس نتیجه می‌گیریم همبستگی ندارند.
- همانند قسمت قبل این سوال را برای قیمت و **number of reviews** بررسی می‌کنیم.
  باز هم مقدار **p value** کمتر از **alpha** است و این دو فیچر هم همبستگی ندارند.

۱.۱.۶- پیش بینی قیمت

برای این کار دوباره تغییراتی در دیتاست به وجود می‌آوریم.

- حذف کردن ستون هایی که متوجه شدیم کارآمد نیستند.
- انکود کردن فیچر های کتگوریکال (**nighbourhood_group , room_type**)

در این مرحله دیتاست آماده داده شدن به مدل است
دیتا ست را به دوبخش تقسیم کرده:
۸۰ درصد ترین
۲۰ درصد تست
و ستون **price** را به عنوان y جدا می‌کنیم.

مدلی که استفاده می‌کنیم مدل **Linear Regression** از کتابخانه **sklearn** است.
و روی دیتای تست به ما اسکور ۵۴ می‌دهد.