به نام خدا

محمدرضا ضیالاری – 97222057

تمرین اول پروژه اول داده کاوی (Airbnb)

استاد درس : دکتر فراهانی

-مقدمه:

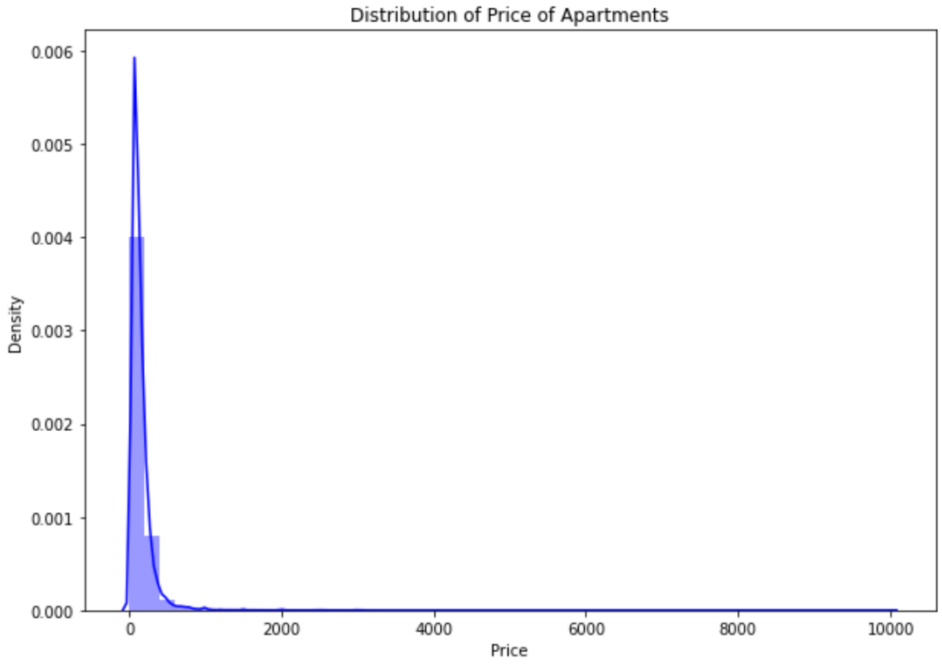
در این تمرین ما با داده های شرکت Airbnb که یک شرکت اجاره مسکن به کسانی است که قصد اقامت موقت در جایی را دارند . در اینجا ما کارهایی آماری را روی این دیتاست انجام می دهیم و نتایج آن را بررسی و تحلیل می کنیم و درنهایت مدلی را برای پیش بینی قیت ارائه می دهیم .

-خواندن داده و پیش پردازش :

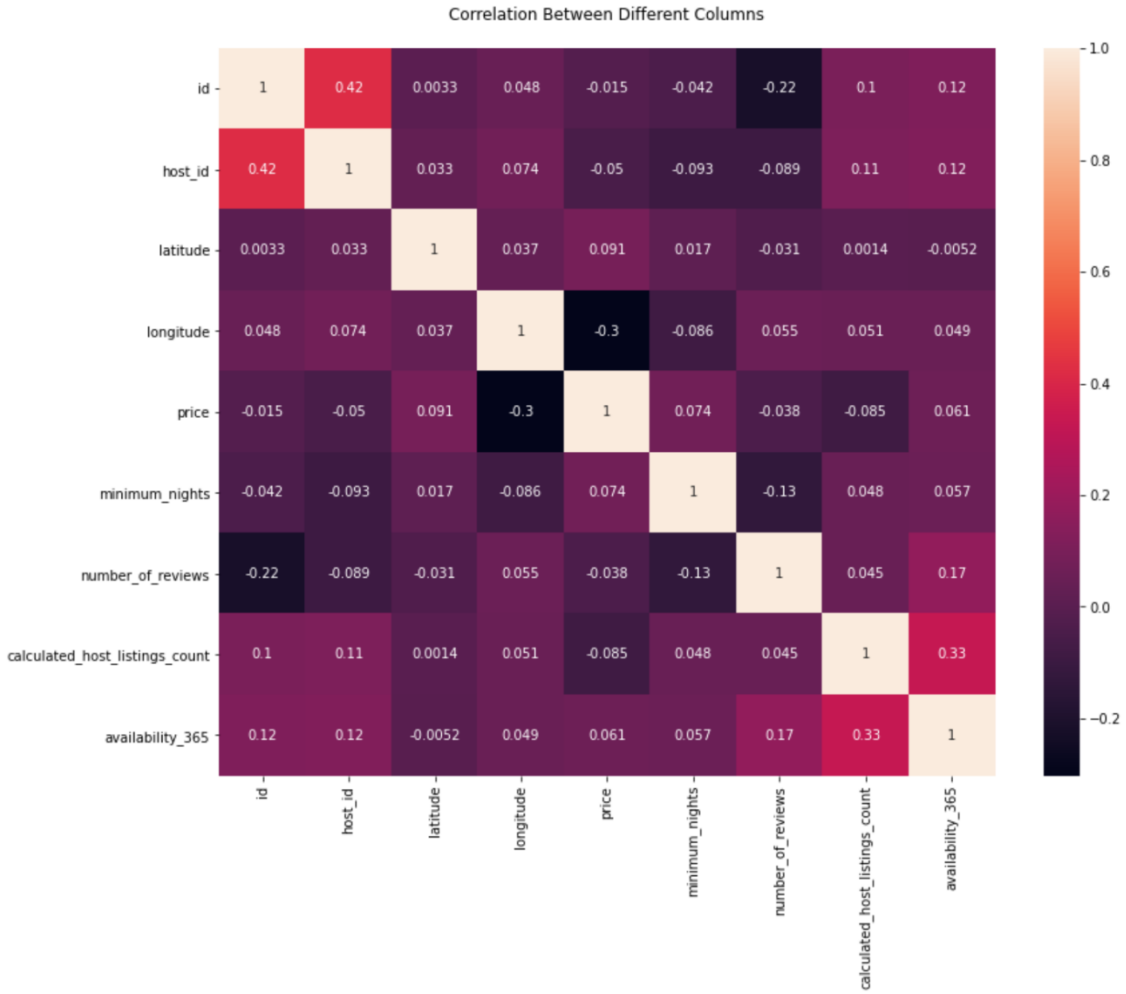
در ابتدا ما داده ها را می خوانیم و ستون هایی که دیتای null زیادی نسبت به بقیه دارند را حذف می کنیم .

-بررسی ها :

یکی از مهم ترین ویژگی هایی که در این دیتاست قرار دارد ، قیمت می باشد . در ابتدا بررسی می کنیم تا ببینیم آیا توزیع قیمت نرمال می باشد یا خیر ؟! که مشاهده می کنیم p_value برابر 0 می شود که بدان معناست که داده های ما نرمال نمی باشد . در زیر شکل توزیع قیمت را می بینیم .

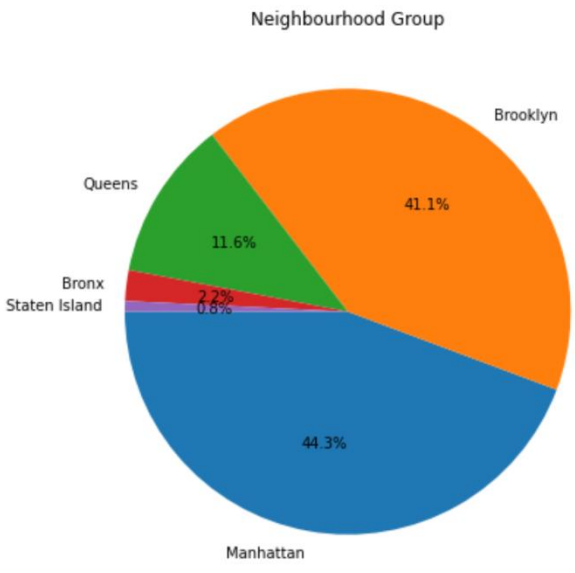


همچنین ماتریس کورولیشن دیتاست را بدست می آوریم تا بررسی کنیم آیا بین هر دو ویژگی(عددی) آیا ارتباط معناداری وجود دارد ؟ که می بینیم بین هیچ دو داده ای(عددی) ارتباط معناداری یافت نمی شود .



- آیا ارتباط منطقی ای بین ترافیک اجاره مسکن و مناطق وجود دارد ؟

داده های ما شامل 5 منطقه می باشند که توزیع آنها بصورت زیر است .



مشاهده می شود که در مناطق بروکلین و منهتن بیشترین خانه برای اجاره وجود دارد .

حال بررسی میکنیم ببینیم کدام مناطق بیشترین ترافیک را داشته اند ؟ برای این کار ما تعداد روز های خالی در طول 365 روز سال برای مناطق مختلف بررسی میکنیم . هر منطقه که تعداد روز های خالی کمتری داشته باشد میتون گفت که ترافیک نسبی آن بیشتر است .

neighbourhood_group	count	mean	std	min	25%	50%	75%	max
Bronx	599.0	158.460768	133.159527	0.0	34.00	132.0	306.50	365.0
Brooklyn	7177.0	94.999443	127.897479	0.0	0.00	14.0	178.00	365.0
Manhattan	2851.0	89.623290	123.950465	0.0	0.00	9.0	170.50	365.0
Queens	2742.0	143.052881	134.700261	0.0	2.00	96.0	281.75	365.0
Staten Island	172.0	200.418605	134.187698	0.0	70.25	230.5	333.25	365.0

مشاهده می شود که دقیقا در بروکلین و منهتن ترافیک بیشتر می باشد پس دلیل بیشتر بودن خانه اجاره ای در این دو منطقه برای مسافران نسبت به دیگر مناطق یک موضوع منطقی است .

- چه کسی بیشترین خانه را دارد ؟ آیا می توان دلیلی منطقی پیدا کرد ؟

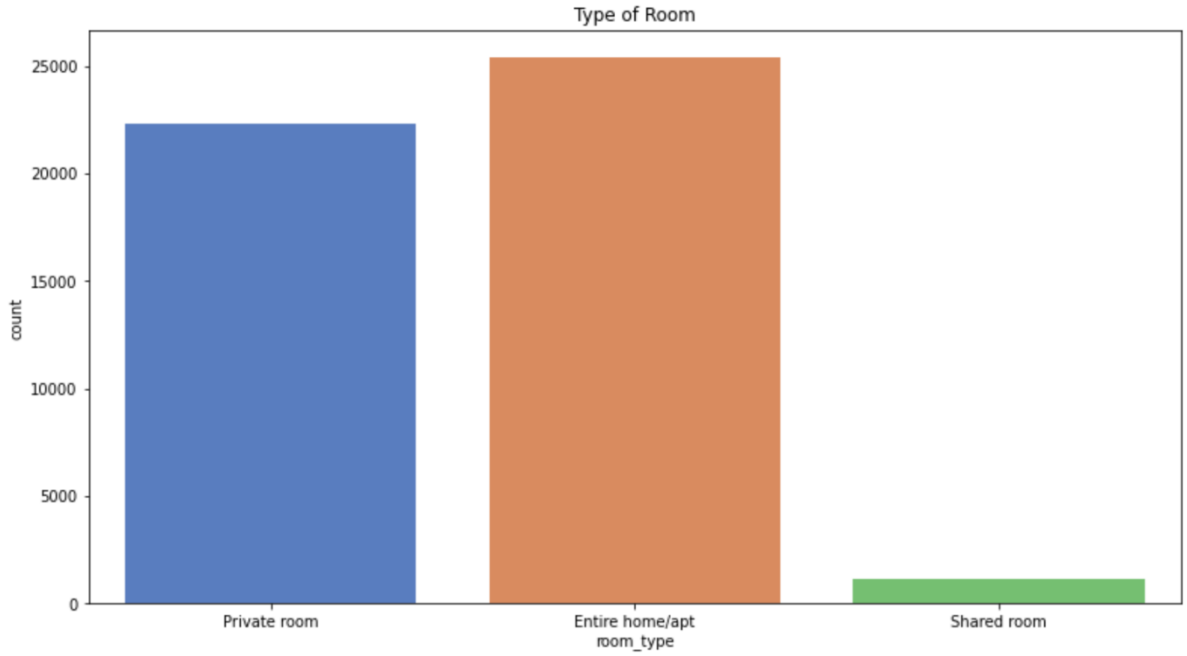
از داده ها می توان استخراج کرد که میزبان با آیدی 219517861 بیشترین خانه را دارد که تمام خانه های او در منهتن هستند . او دارای 327 خانه که شامل 8 عدد private room و مابقی entire home/apt می باشد . نکته جالب اینجاست که طبق اطلاعات زیر به طور میانگین خانه های او در بیشتر روزهای سال خالی هستند و نمیتوانیم دلیلی منطقی تر جز ثروتنمد بودن فرد برای این تعداد خانه پیدا کنیم .

```
count    327.000000
mean     301.492355
std       66.677544
min       37.000000
25%      282.500000
50%      328.000000
75%      341.500000
max      365.000000
```

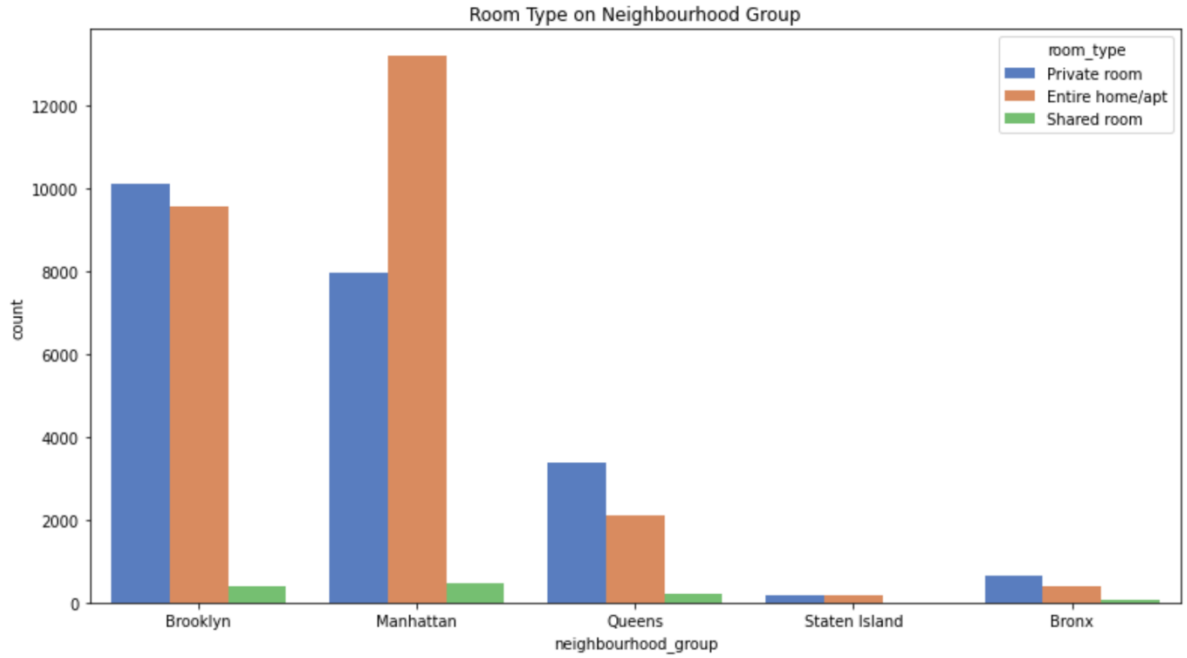
همچنین میانگین قیمت خانه های او (253) از میانگین قیمت کل خانه ها (152.7) و میانگین خانه های منهتن (117) بیشتر است که میتواند دلیلی بر خالی ماندن خانه ی او درطول سال باشد .

-بررسی انواع خانه ها:

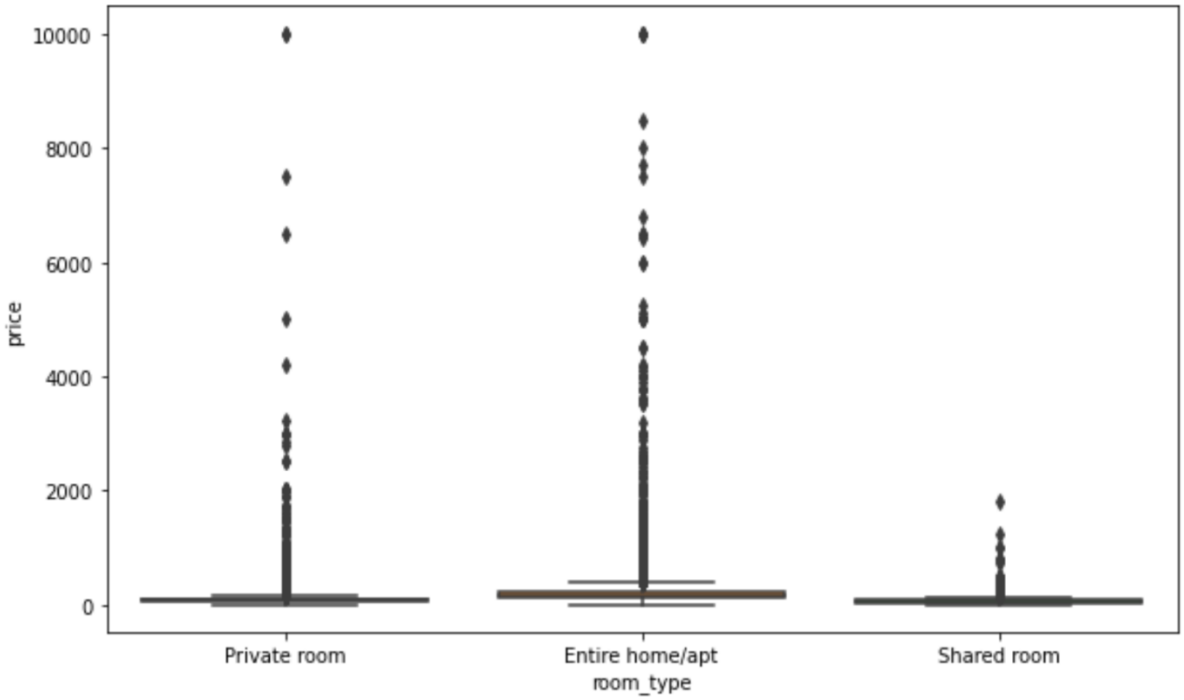
در این دیتاست ما سه نوع خانه private room و entire home/apt و shared room داریم که نحوه توزیع آن روی کل داده ها بصورت زیر است .



و برای هر منطقه به تفکیک بصورت زیر می باشد :

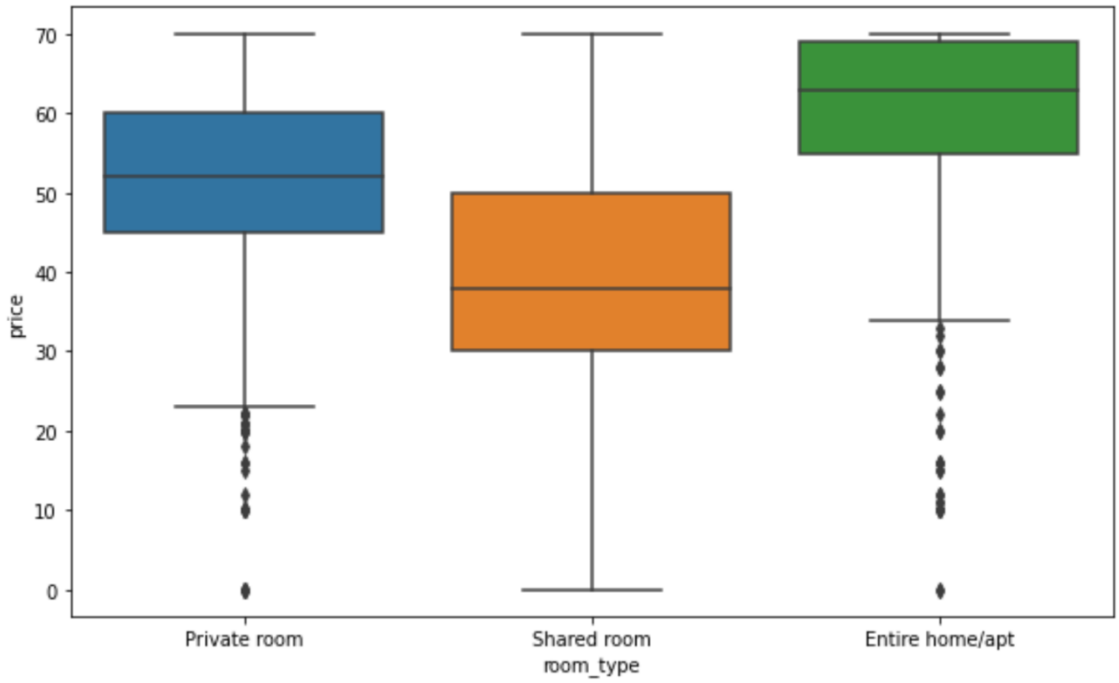


سپس دامه تغییرات قیمت هر کدام از این نوع انه ها را بررسی کردیم و بصورت یک نمودار جعبه ای رسم نمودیم .



که می بینیم دامنه تغییرات قیمت بسیار زیاد است . دامنه تغییرات را کوچکتر و منطقی تر میکنیم و داده های پرت را از دیتاست حذف می کنیم و نمودار جعبه ای بهصورت زیر می شود .

*نکته ک بعد از حذف داده های پرت نیز بررسی کردیم که ببینیم آیا توزیع قیمت نرمال شده است یا خیر ؟ اما همچنان p_value برابر 0 بود که به معنی عدم توزیع نرمال قیمت می باشد .



room_type	count	mean	std	min	25%	50%	75%	max
Entire home/apt	851.0	59.473561	11.872467	0.0	55.0	63.0	69.0	70.0
Private room	11839.0	52.665597	11.445542	0.0	45.0	52.0	60.0	70.0
Shared room	851.0	40.828437	14.152154	0.0	30.0	38.0	50.0	70.0

طبق جدول بالا میبینیم که میانگین قیمت اتاق های اشتراکی کمتر است که توجیه منطقی نیز دارد .

-آزمون های فرض:

در انتهایا این تمرین آزمون های فرضی را برروی ویژگی های مختلف انجام دادیم که نتایج آن بصورت زیر است:

1- میانگین قیمت در نوع خانه های مختلف با یکدیگر اختلاف معناداری دارند و p_value کوچکتر از آلفا(0.05) بود .(oneway ANOVA)
2- میانگین قیمت بین مناطق مختلف نیز بایکدیگر اختلاف معناداری داشتند و p_value کوچکتر از آلفا(0.05) بود . .(oneway ANOVA)
3- توزیع نوع خانه ها در مناطق مختلف با یکدیگر اختلاف معناداری دارند و p_value کوچکتر از آلفا(0.05) بود .(خی-دو)
4- میانگین قیمت بین میزبان های مختلف باهم اختلاف معناداری داشتند و کوچکتر از آلفا(0.05) بود .
5- میانگین قیمت خانه های پرطرفدار و شلوغ(در طول سال) به میانگین قیمت کل خانه ها نزدیک بود که می توان نتیجه گرفت خانه هایی با قیمت متوسط پرطرفدارتر هستند .

-مدل پیش بینی قیمت:

ابتدا اطلاعاتی از قبیل نام میزبان و آیدی را حذف میکنیم . سپس اطلاعات مربوط به همسایه ها به دلیل تنوع زیاد و همچنین آیدی میزبان به دلیل دامنه تغییرات زیاد را حذف میکنیم .

سپس نوع اتاق و گروه همسایگی را که داده های استرینگ هستند وان هات میکنیم .

80 درصد داده ها را به عنوان داده آموزشی و 20 درصد را به عنوان داده تست جدا می کنیم .

در اینجا برای پیش بینی قیمت از رگرسیون لاجستیک استفاده کردیم . پس از فیت شدن مدل ، داده های تست را به مدل میدهیم تا فرایند پیش بینی را طی کند و سپس دقت مدل را با تلرانس 3 دلار اختلاف قیمت بررسی میکنیم .

مشاهده می شود که دقت مدل حدود 74.45 درصد می باشد که دقت مناسبی می باشد .