

گزارش تمرین شماره 1

واحد درسی داده کاوی

اساتید محترم :

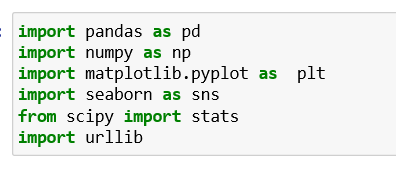
جناب آقای دکتر فراهانی

جناب آقای دکتر خردپیشه

هدیه آشوری 99422022

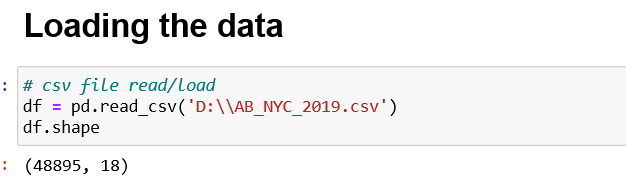
15/01/1400

* بررسی داده ها



در ابتدا پکیج های مورد نیاز را Import

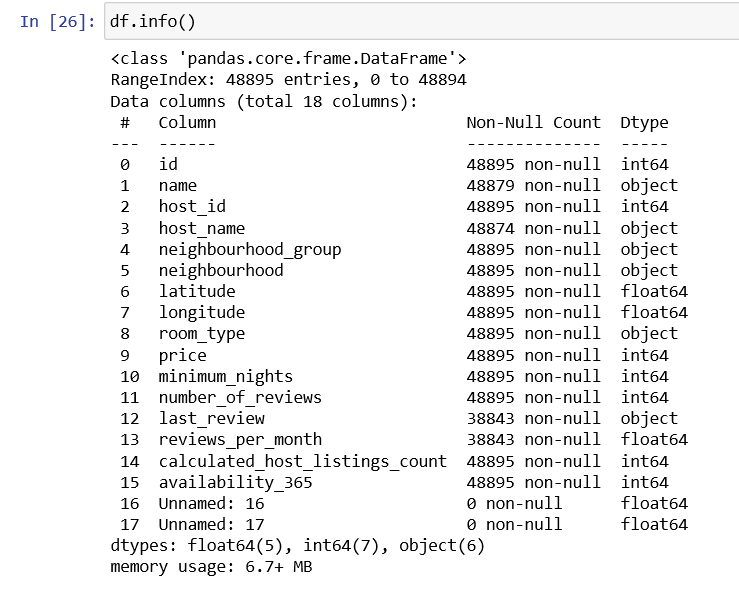
می کنیم :

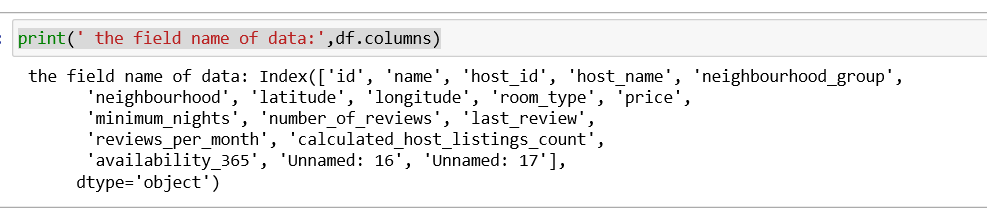


سپس data مورد نظر را با دستور ذیل فراخوانی می کنیم

و

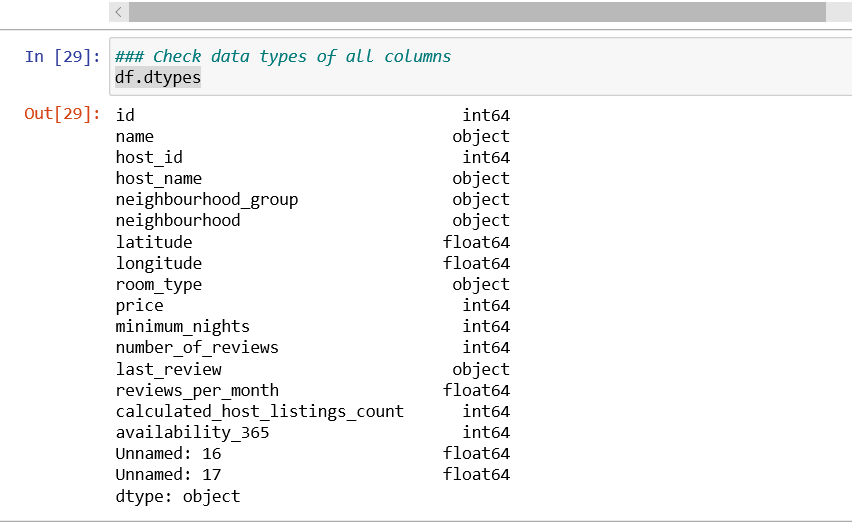
با دستور df.shape ابعاد DataFrame را به دست می آوریم ( 18 ستون و 48895 ردیف )

با دستور df.info() خلاصه مختصری از داده ها را به شرح ذیل دریافت می کنیم .

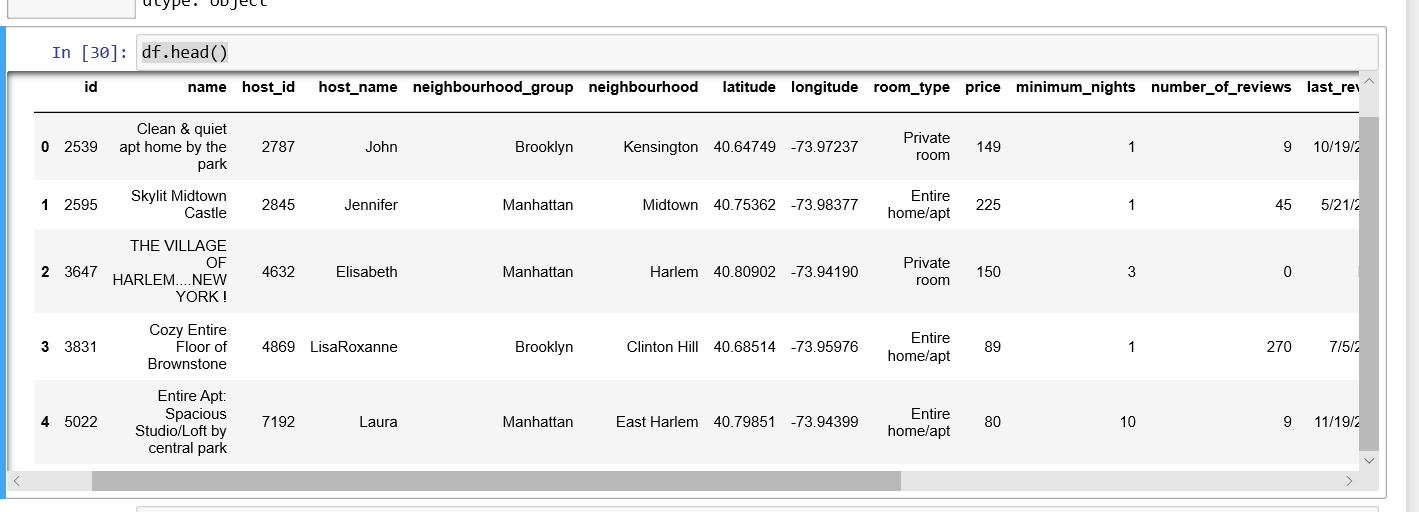
با دستور print(' the field name of data:',df.columns) نام ستون ها را دریافت کردم

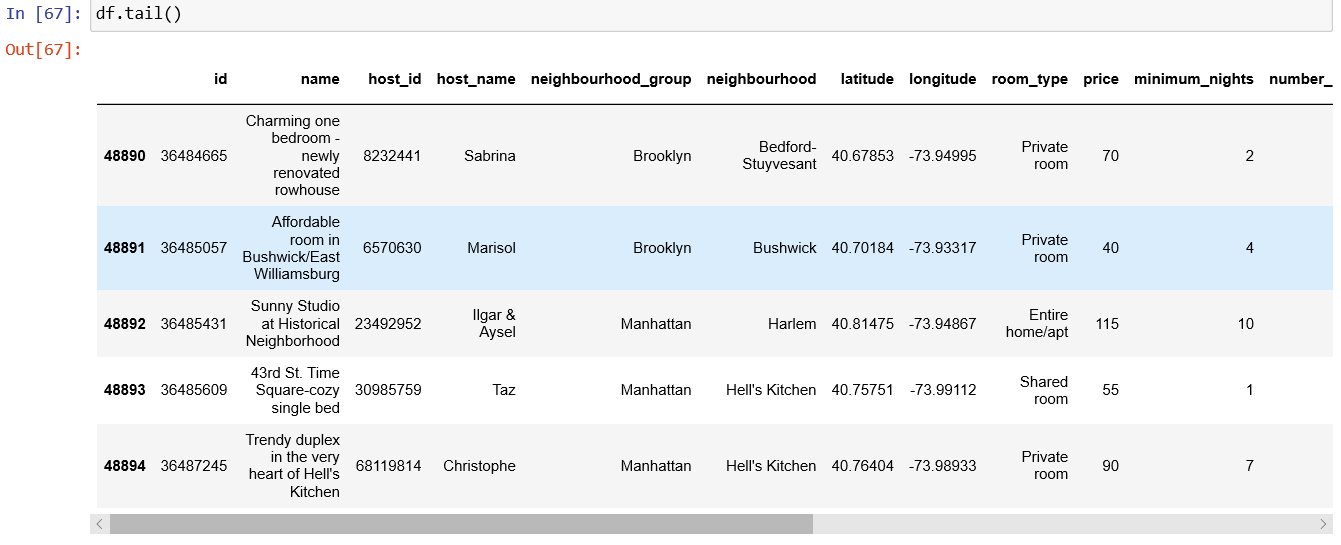
با دستور df جدولی از داده دریافت کردم



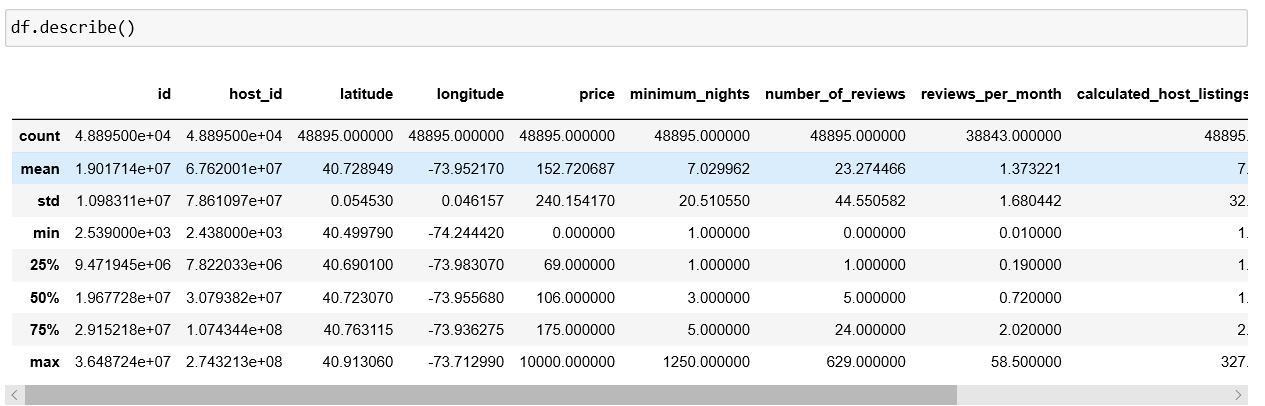
با دستور df.dtypes ، نوع دیتا ها را بررسی کردم

با دستور df.head() ، پنج خط اول داده ها را دریافت کردم

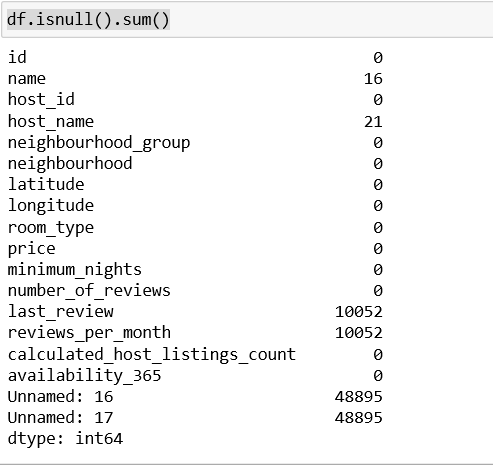


با دستور df.tail() پنج خط آخر داده ها را مشاهده کردم

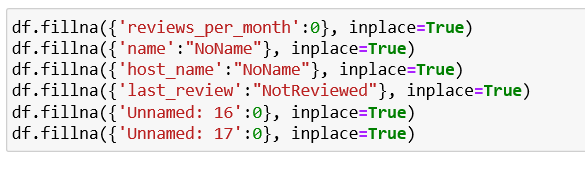
با دستور df.describe () خلاصه ای از اطلاعات عددی data را به دست آوردم



با دستور df.isnull().sum() داده های null را بررسی کردم

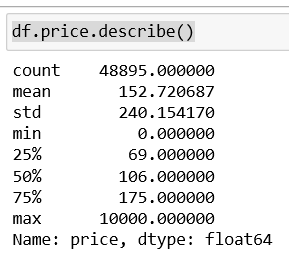


با دستور ذیل ، داده های null را پر کردم





مجددا با دستور df.isnull().sum() به بررسی نتیجه پرداختم   
همانگونه که می بینیم تمامی صفر شده است

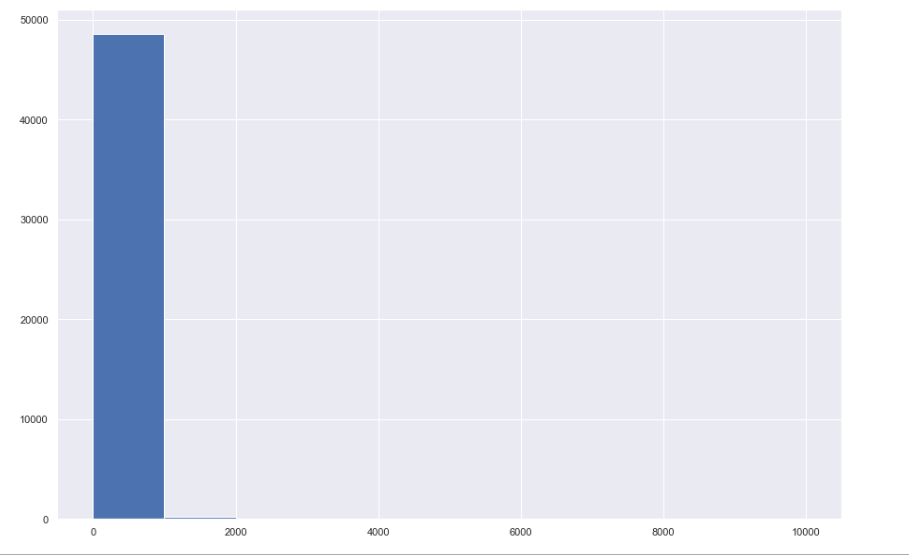


با df.price.describe() ستون قیمت را بررسی کردم

در این جا مشاهده می کنیم که میانگین صفر هست

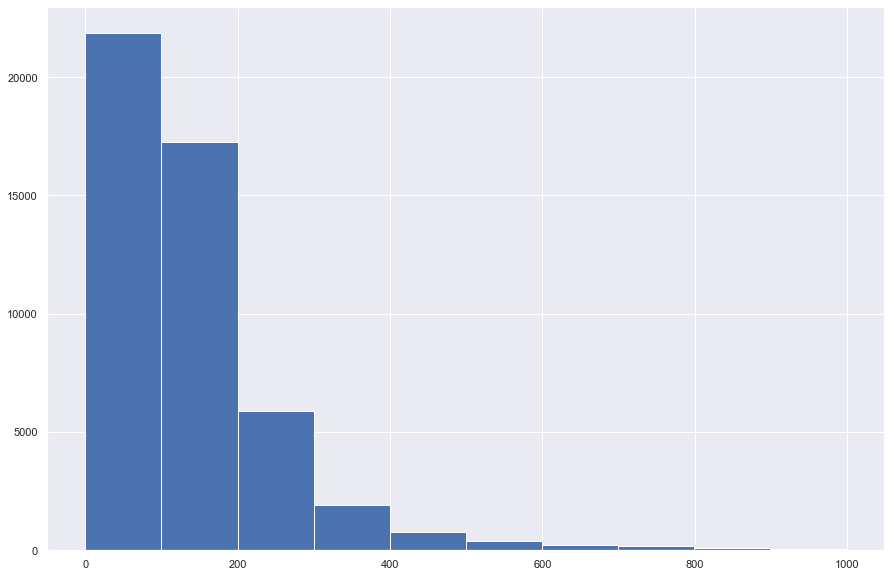
**توزیع قیمت را مشاهده خواهیم کرد**

hist\_price=df["price"].hist()



در نمودار دریافت شده می بینیم که قیمت اکثر لیست ها کمتر از 1000 دلار هست .

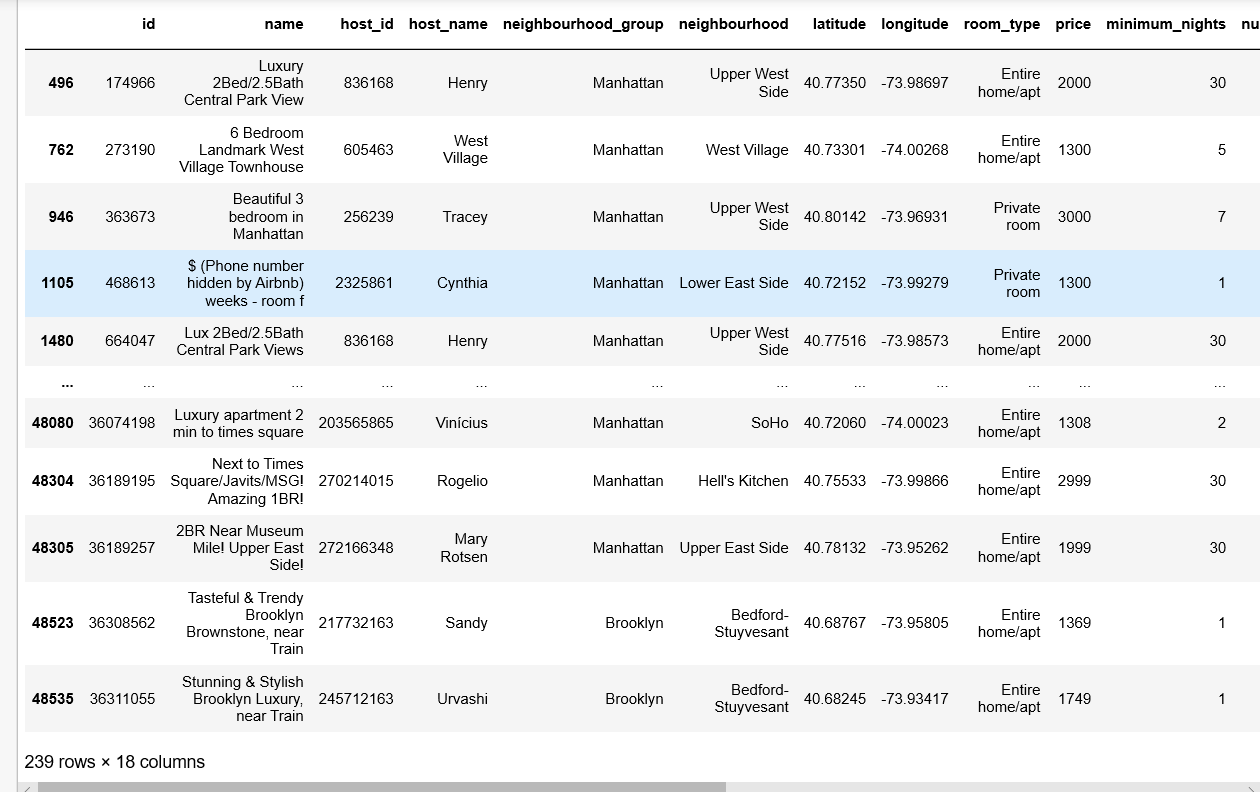
با دستور hist\_price1=df["price"][df["price"]<1000].hist() نمودار ذیل را داریم و اجازه دادیم که هیستوگرام با قیمت های کمتر از 1000 دلار رسم گردد



تصویر واضح تری داریم اما خوب می بینیم که توزیع به سمت چپ متمایل هست و چولگی دارد

حالا می خواهیم ببینیم که چه تعدادی از لیست ما مبلغ بیشتر از 1000 دلار داشته اند؟

با دستور dataset =df[df["price"]>1000] این مورد را انجام داده و با دستور dataset جدول ذیل را دریافت می کنیم که دارای 239 ردیف هست

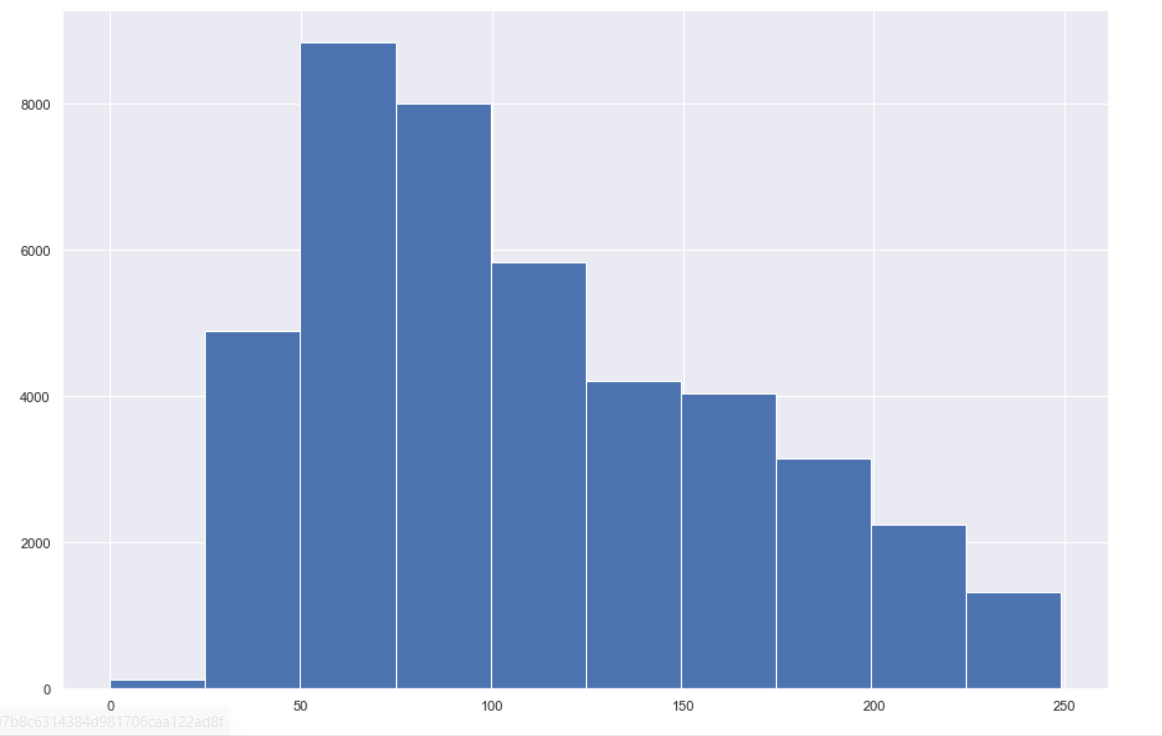


این جدول می گوید که 239 تا از لیست ما دارای قیمت روزانه بالاتر از 1000 دلار می باشد این لیست بسیار باشکوه و رویایی است اما مطمعنا در زمان ورود خطایی رخ داده است

با دستور dataset=df[df["price"]<1000] قیمت های کمتر از 1000 را در dataset ریختیم

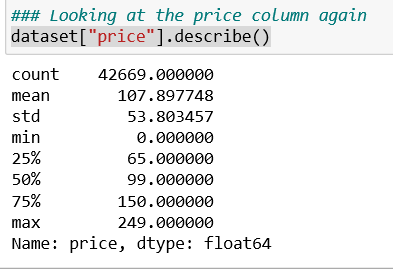
حالا می خواهیم نموداری از قیمت های بالاتر از 250 دلار رسم کنیم :

hist\_price2=dataset["price"][dataset["price"]<250].hist()



در این جا ما نمودار توزیع گاوسی بهتری داریم   
پس آستانه قیمت را روی 250 قرار می دهیم

مجددا با دستور dataset["price"].describe() ستون قیمت را بررسی می کنیم

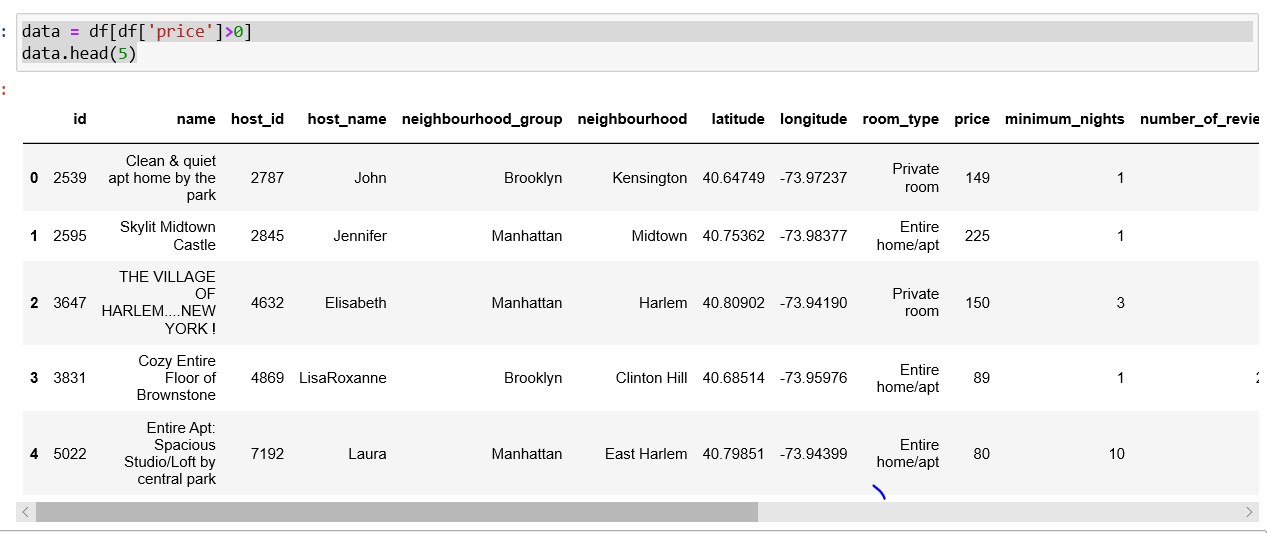


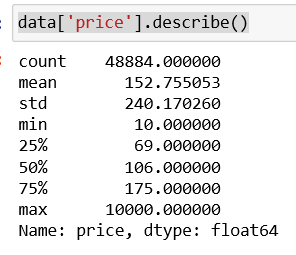
همانگونه که مشاهده می کنید در این جا میزان min قیمت صفر هست   
سوالی که پیش می آید این است که آیا ممکن هست ما خانه ای را اجاره کنیم و هزینه ای پرداخت نکنیم ؟!!   
طبیعتا خیر

پس data من مشکل دارد و من باید آن را اصلاح کنم

بنابراین اقدام به data cleaning می کنم و مقادیر price=0 را حذف می کنم

با دستور data = df[df['price']>0] و data.head(5) داده های price=0 را حذف کردم و 5 ردیف اول نمایش داده شد





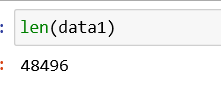
مجددا با دستور data['price'].describe() اطلاعات قیمت را بررسی می کنیم همانطور که مشاهده می کنید min با مقدار صفر حذف شد

حالا با دستور len طول داده قیمت را به دست می آوریم

می خواهیم داده های outlier را در ستون قیمت یا price پیدا کنیم

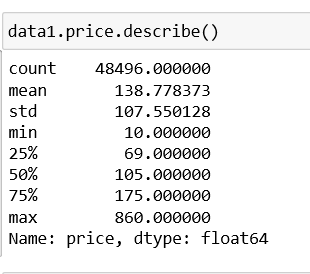
با دستور data1 = data[data['price']<= data['price'].mean() + 3\*data['price'].std()]

و رنج 3- برابر std و 3 برابر std را حذف می کنیم

مجددا با دستور len طول داده قیمت را به دست می آوریم

همانگونه که مشاهده می کنید تعدادی داده حذف شده است و

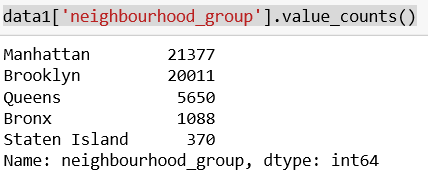
طول داده کاهش یافته است



مجددا با دستور data['price'].describe() اطلاعات قیمت را بررسی می کنیم مشاهده می کنید که مقدار std , … تغییر کرده است

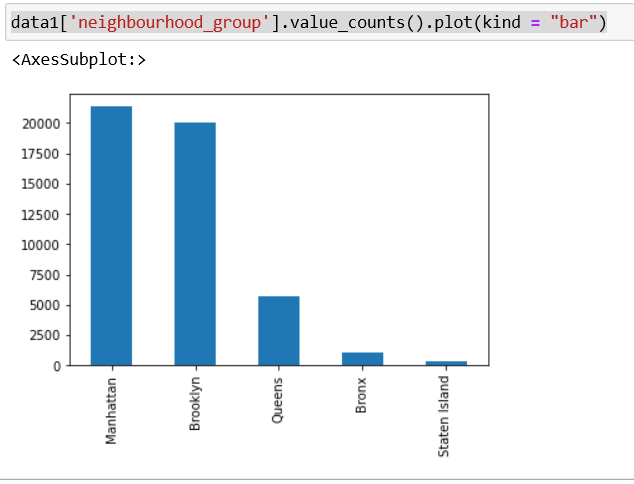
حالا می خواهیم ببینیم در نواحی پنج گانه نیویورک چقدر آگهی داشته ام با دستور ذیل این مورد را انجام می دهیم

data1['neighbourhood\_group'].value\_counts()



با دستور ذیل برای رسم نمودار اقدام می کنیم :

data1['neighbourhood\_group'].value\_counts().plot(kind = "bar")



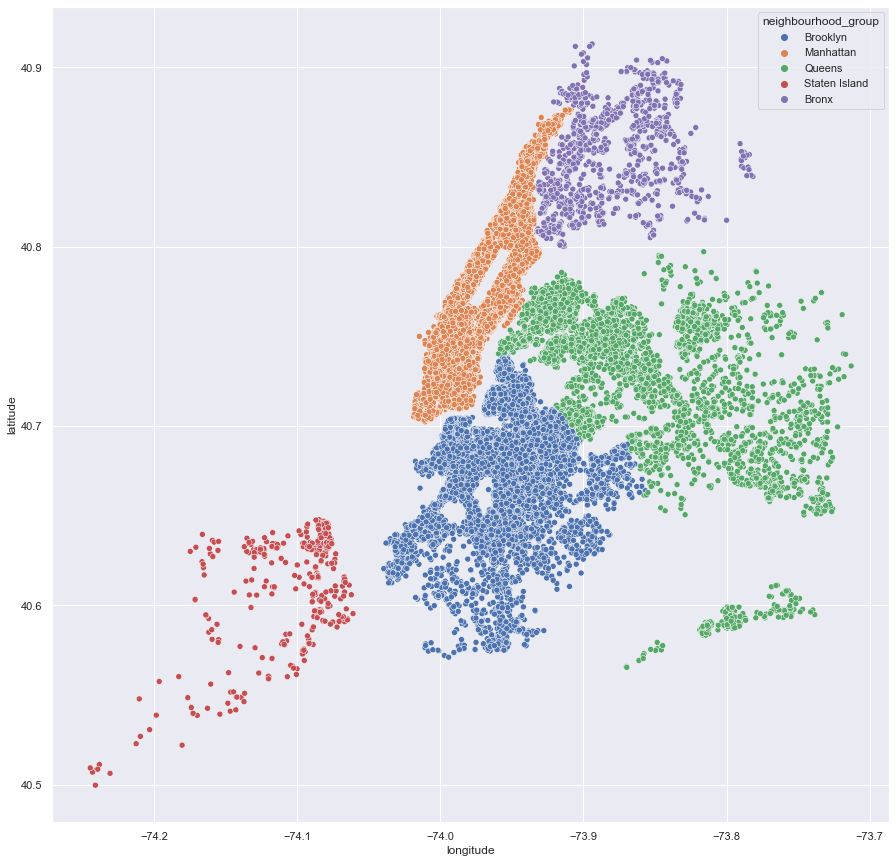
به ترتیب از چپ به راست بیشترین آگهی تا کمترین آگهی نمایش داده شده است .

Manhattan بیشترین Request را داشته است و Staten Island کمترین Request را داشته است

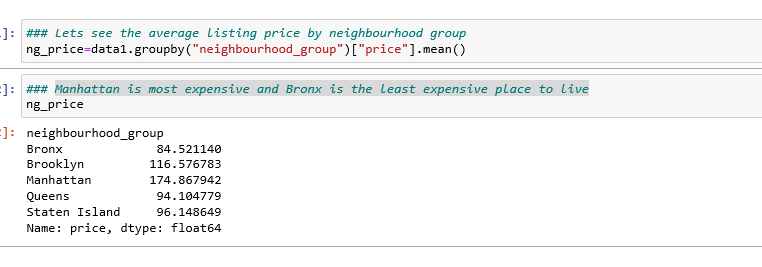
حال می خواهم از طریق scatterplot پراکندگی request ها را در نیویورک نمایش دهم

plt.figure(figsize=(15, 15))

sns.scatterplot(x=data.longitude,y=data.latitude,hue=data.neighbourhood\_grou)



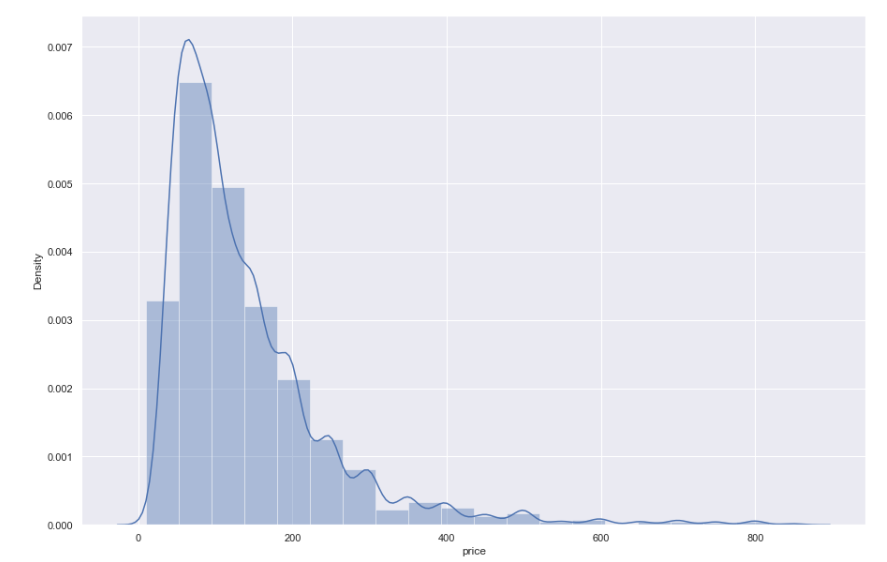
گرانترین و ارزان ترین شهر در نیویورک ؟

منهتن گران ترین و برانکس کم هزینه ترین مکان برای زندگی است

Distribution data

توزیع قیمت را بررسی کنیم :

sns.set(rc={'figure.figsize':(15,10)})

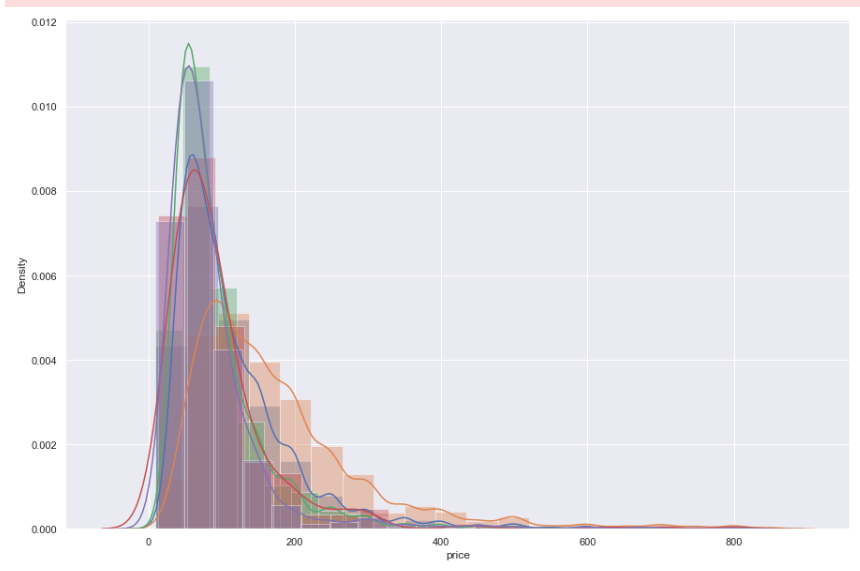
sns.distplot(data1['price'],kde\_kws={"label": 'price'}, bins=20)

همانطور که در شکل می بینیم توزیع نرمال نیست   
و باید روی این data تغییرات بدهیم تا دیتا نرمال گردد

توزیع را برای هر همسایه هم بررسی می کنیم :

sns.set(rc={'figure.figsize':(15,10)})

for groups in data1.neighbourhood\_group.unique():

 sns.distplot(data1.price[data1['neighbourhood\_group']==groups],kde\_kws={"label": groups}, bins=20)

همانطور که مشاهده می کنید

اینجا هم توزیع نرمال نیست و

ما چولگی داریم

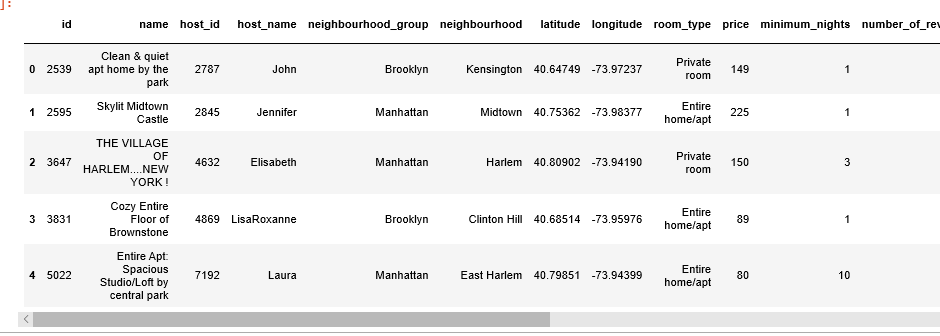
برای اینکه بتوانیم data را نزدیک به نرمال کنیم از transfer ها استفاده می کنیم

به طورمثال می توان از price، loge  گرفت . بنابراین یک فیچر جدیدی تعریف می کنیم به نام price\_log\_e که همان price ما هست و از آن log گرفتیم در پایه e

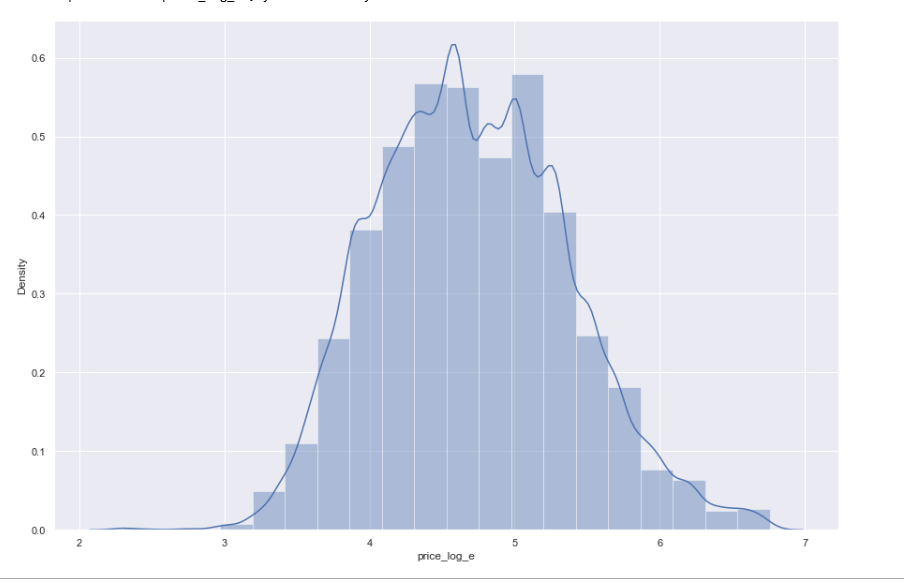
data1['price\_log\_e'] = np.log(data1['price'])

data1.head(5)

با اعمال کد فوق ، 5 ردیف اول به شرح ذیل می باشد :



sns.set(rc={'figure.figsize':(15,10)})

sns.distplot(data1['price\_log\_e'],kde\_kws={"label": 'price in log e'}, bins=20)

همانطور که مشاهده می کنیم تقریبا دیتا نرمال سازی شد

از کجا مطمئن شوم که نرمال سازی شده است ؟با آزمون فرض -- Normal test

stats.normaltest(data1["price\_log\_e"])

NormaltestResult(statistic=562.3974954306634, pvalue=7.532437713364285e-123)

میزان pvalue بسیار کم است ( نزدیک به صفر هست ) و اینطوری فرض نرمال سازی ما رد می شود اما برای ادامه با مقداری خطا آن را نرمال فرض می کنیم .

برای نرمال سازی بهتر از log10 استفاده می کنیم و یک فیچر جدید تحت عنوان price\_ log\_10 ایجاد می کنیم و

data1['price\_log\_10'] = np.log10(data1['price'])

data1.head(5)

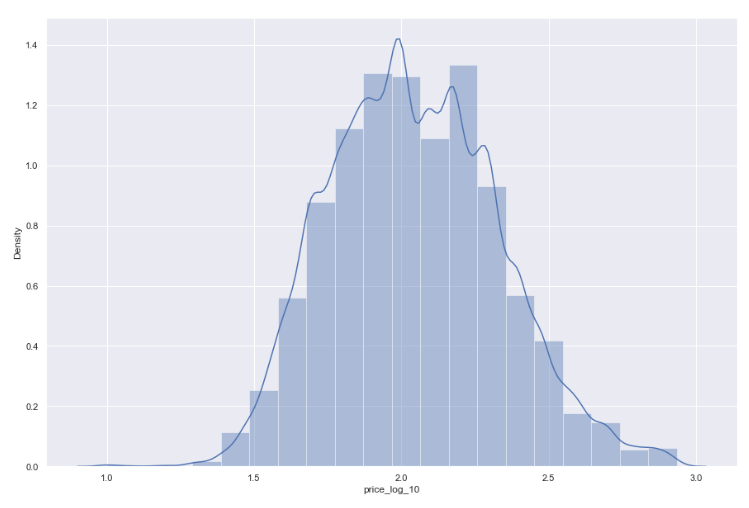
پیرو این دستور 5 ردیف اول به شرح ذیل خواهد بود.



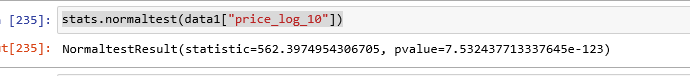
از طریق دستور seaborn نمودار می کشیم

sns.set(rc={'figure.figsize':(15,10)})

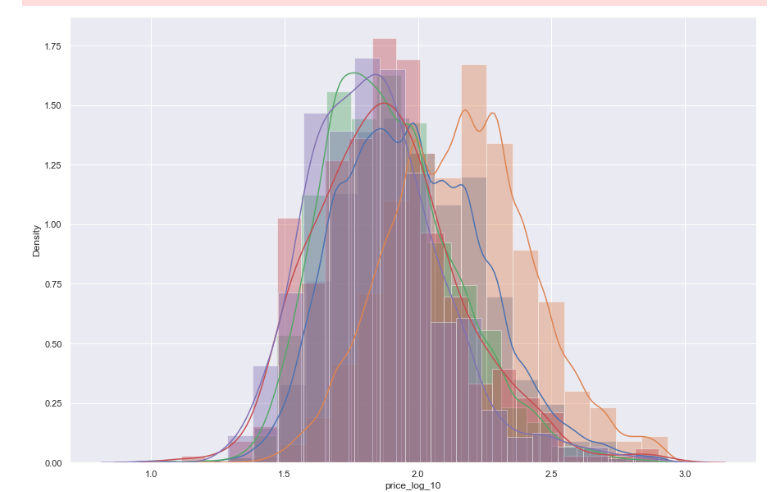
sns.distplot(data1['price\_log\_10'],kde\_kws={"label": 'price in log 10'}, bins=20)



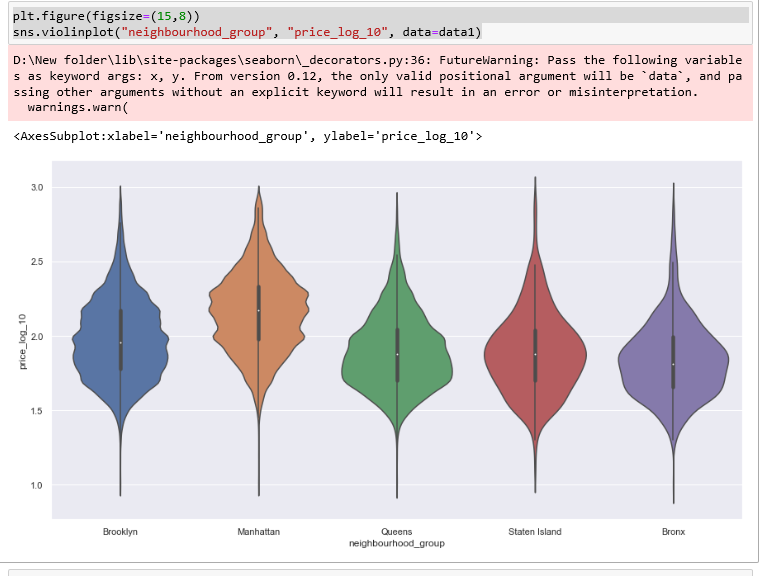
آزمون نرمال تست : با توجه به میزان pvalue ، price\_ log\_10 به دیتای نرمال نزدیک تر هست



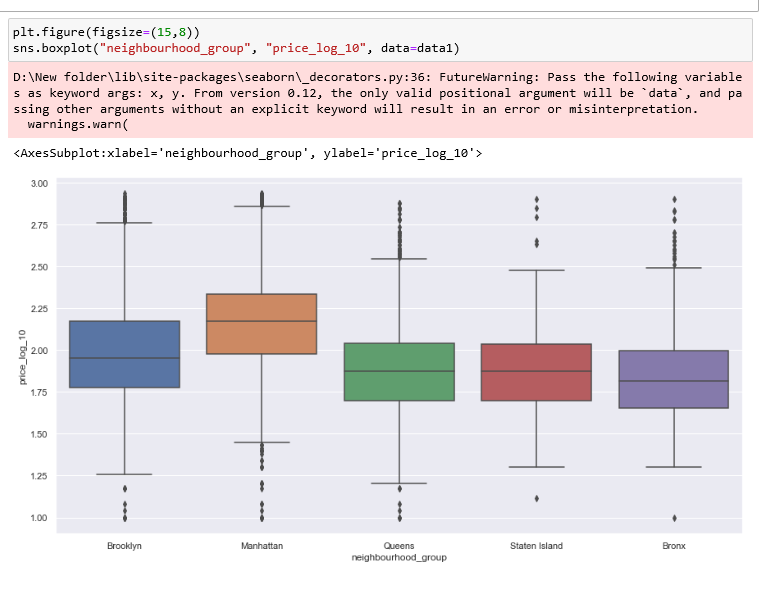
دیتاهای مختلف را اینجا plot می کنیم



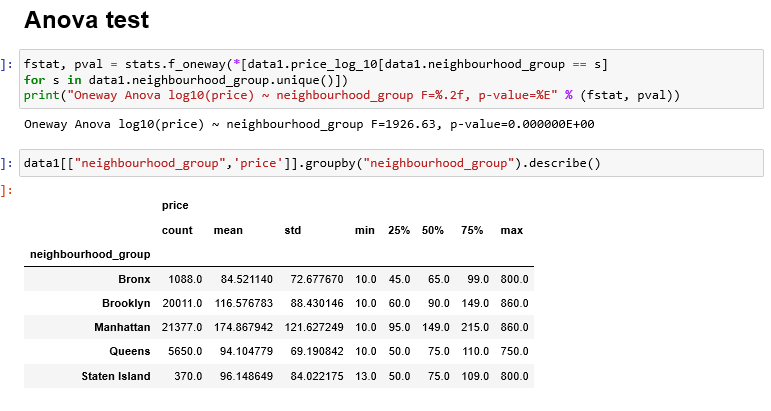
آیا رابطه ای بین price\_ log\_10 و مناطق پنج گانه هست ؟



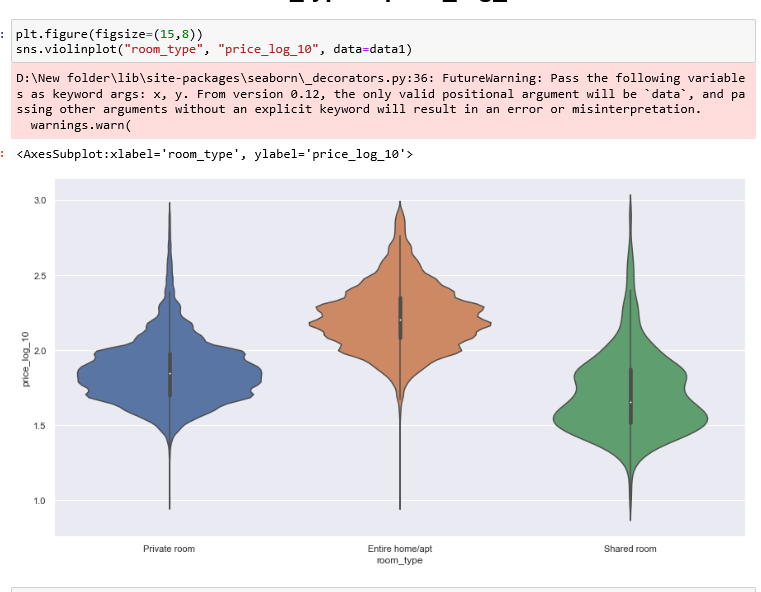
نمودار violinplot می کشیم . نموداریکه در آن mean, max مشخص هست

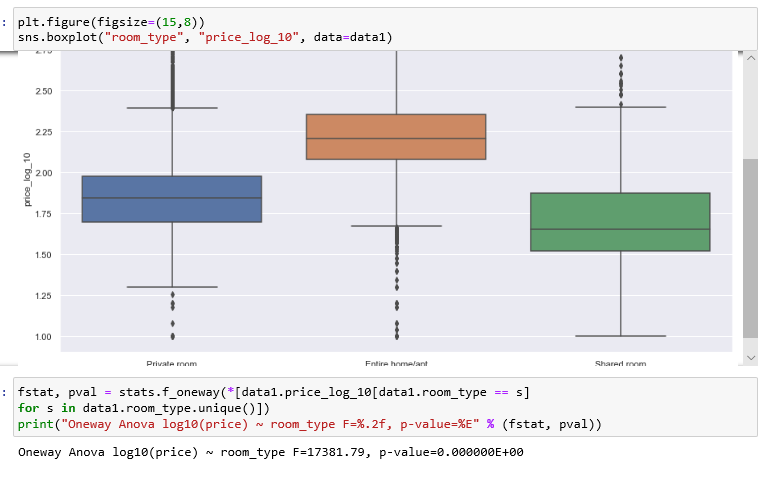


نمودار plotbox کشیده ایم و سپس از anovatestاستفاده می کنیم



باتوجه به اینکه pvalue صفر شده است فرض خلف ما رد شده و فرض اصلی تایید می گردد . پس رابطه ای وجود دارد و منهتن بیشترین درآمد را داشته است

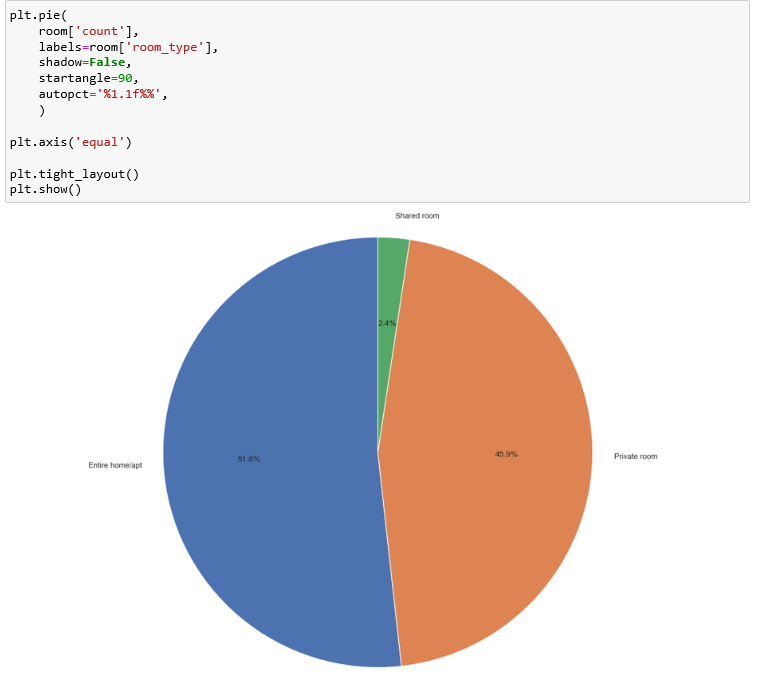
آیا رابطه ای بین نوع خانه و قیمت هست ؟

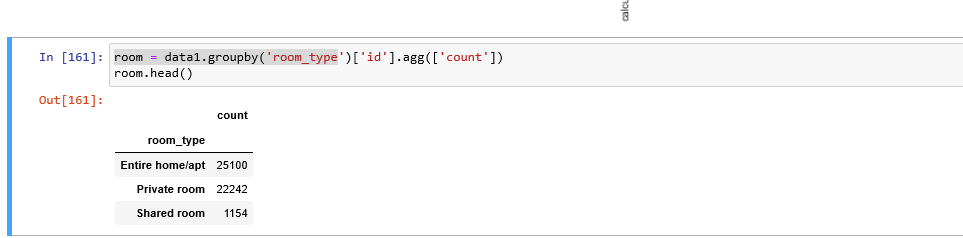


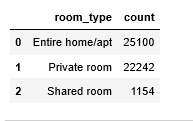
باتوجه به اینکه pvalue صفر شده است فرض خلف ما رد شده و فرض اصلی تایید می گردد . پس رابطه ای وجود دارد براساس نوع خانه مبلغ متفاوت هست خانه مستقل گرانترین و اتاق خصوصی میانه و اتاق اشتراکی کمترین هزینه را دارد

فراوانی اتاق ها را محاسبه کنید؟

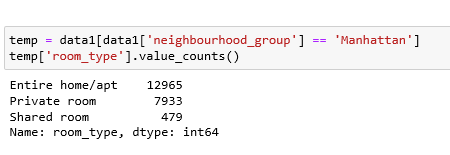
نمودار خانه ها :

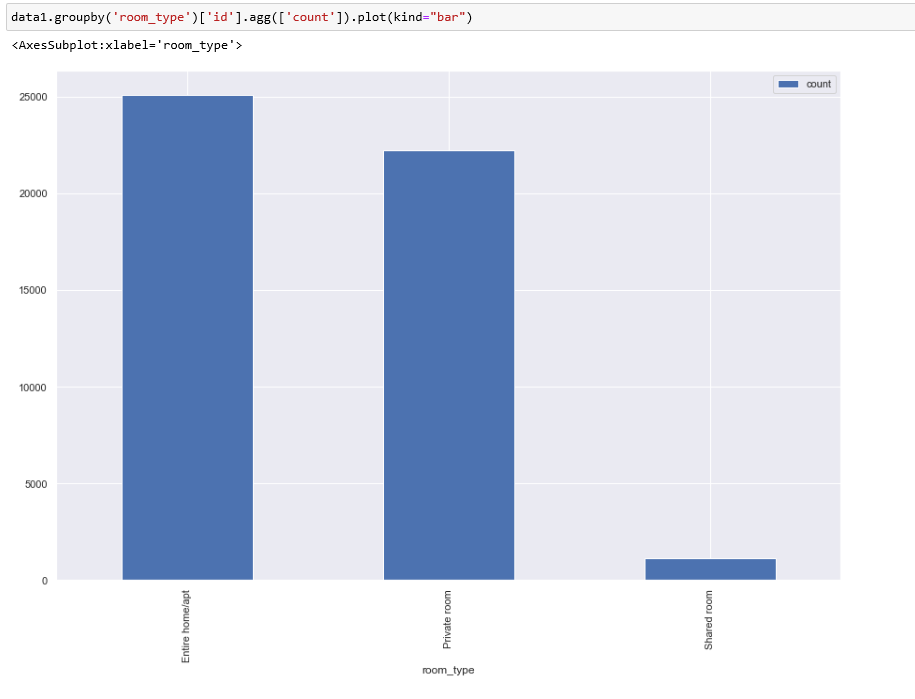




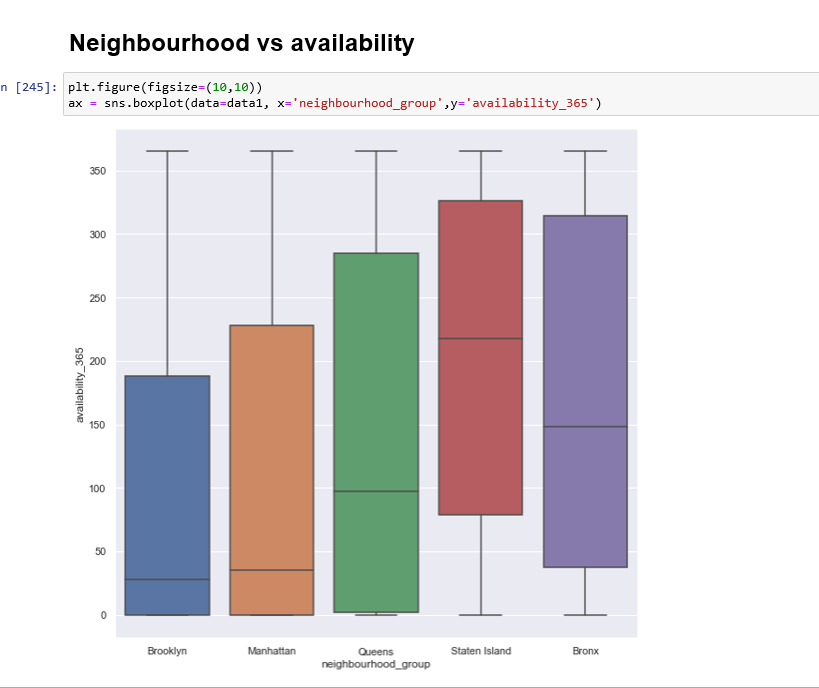
 چند نوع خانه داریم ؟   
همان طور که مشخص است سه نوع   
آپارتمانی یا سوئیت   
خصوصی   
عمومی

* تعدادخانه ها ؟
* تعداد خانه در manhatan چند تا هست ؟



تعداد خانه ها بر اساس host\_id مشخص کنید ؟

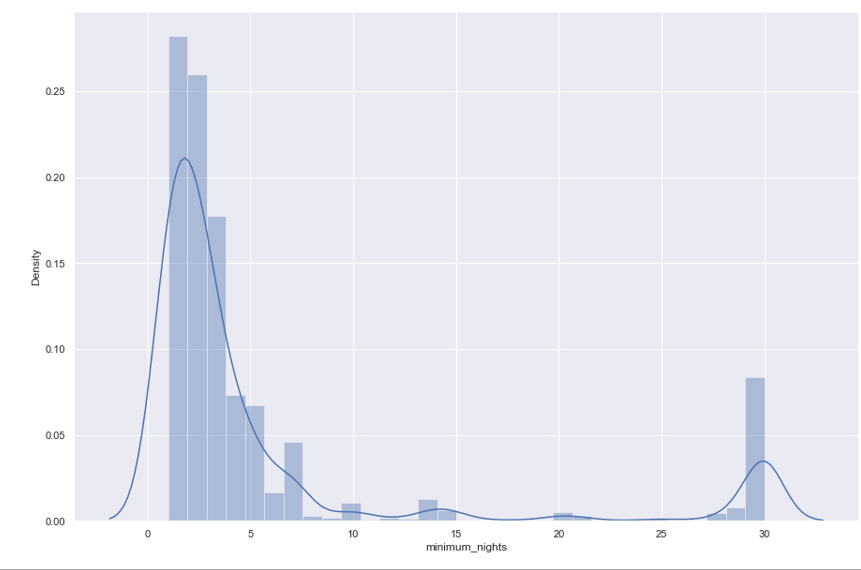


جالبه که اتاق اشتراکی در طول سال بیشترین میزان را دارد .

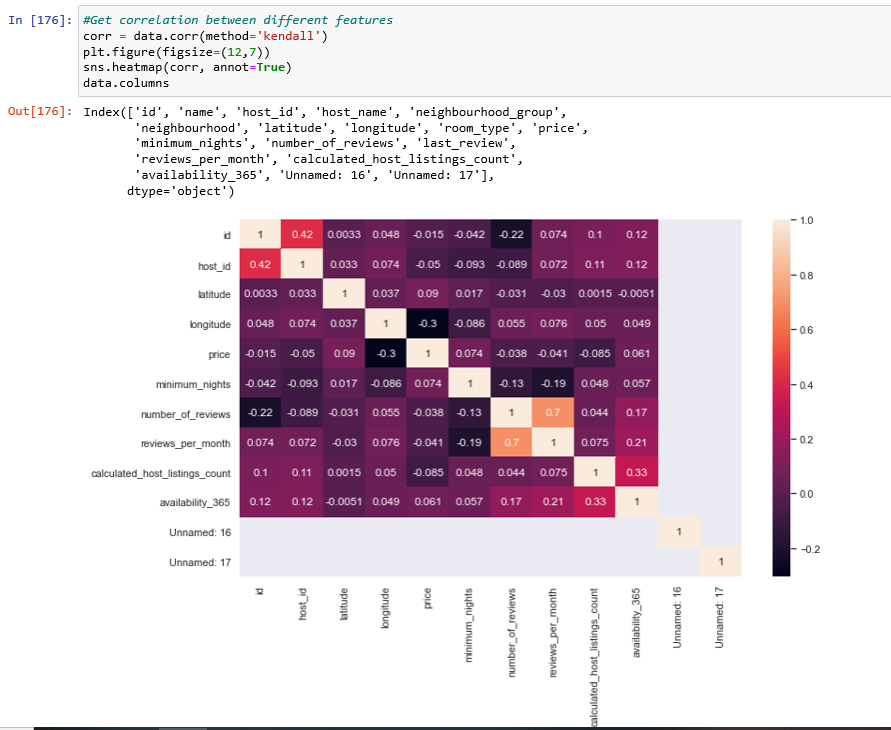
شب های رزرو شده به چه صورت است ؟

sns.distplot(df[(df['minimum\_nights'] <= 30) & (df['minimum\_nights'] > 0)]['minimum\_nights'], bins=31)

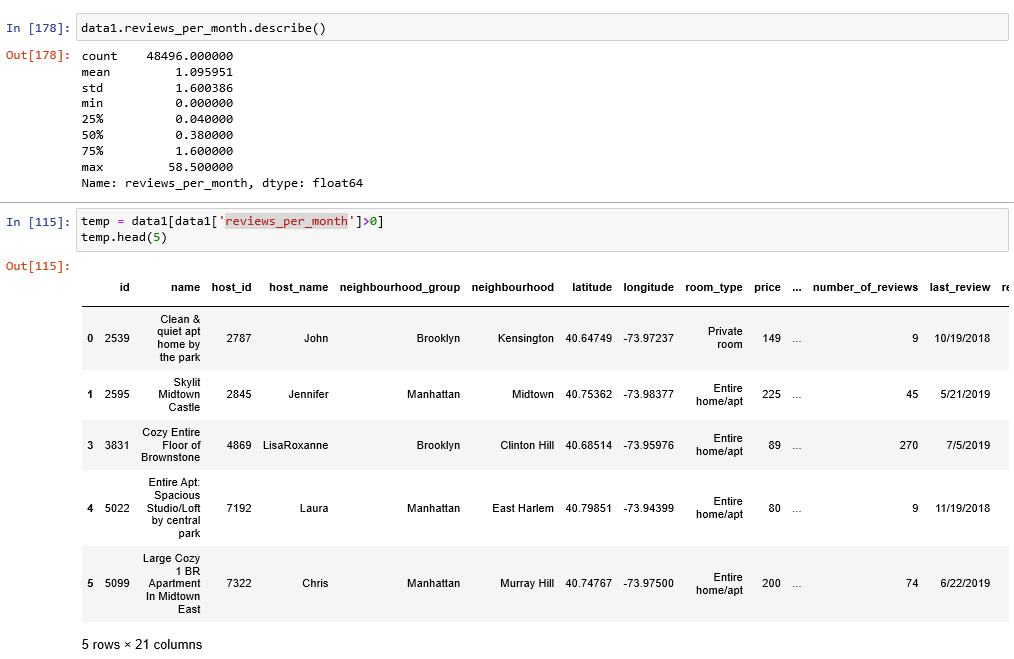
plt.ioff()



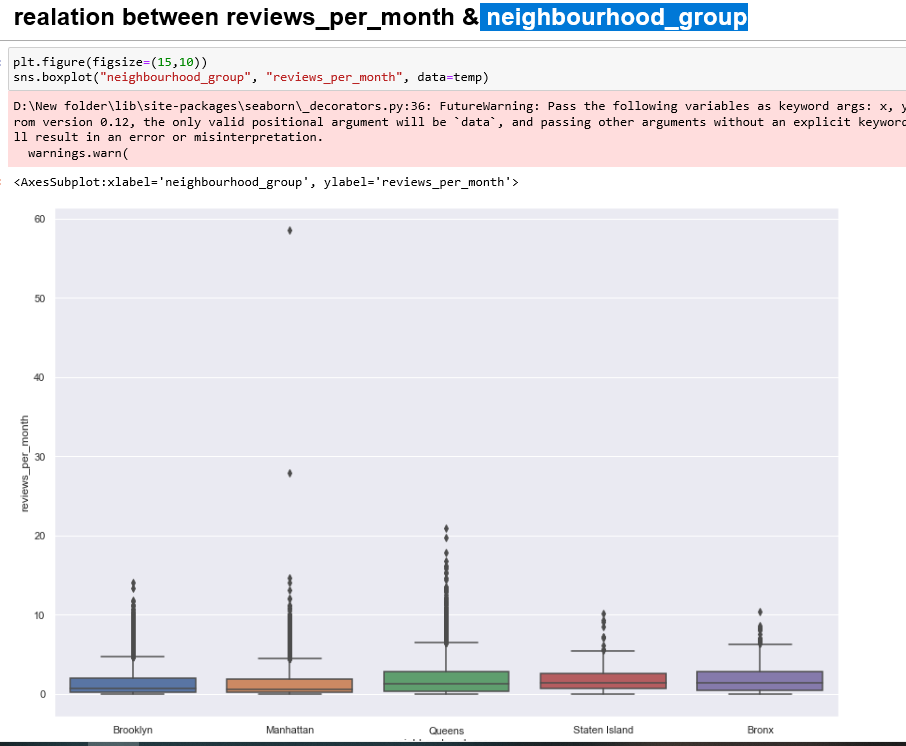
Correlation بین فیچرهای مختلف

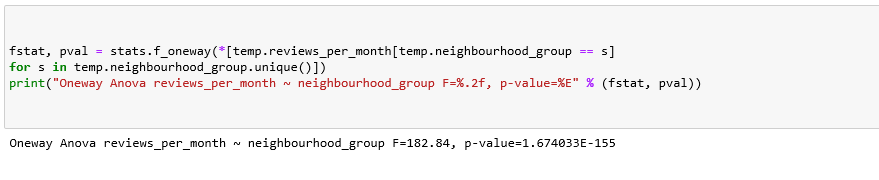


• What can we learn from predictions? (ex: locations, prices, reviews, etc)

بررسی اطلاعات reviews\_per\_month

آیا رابطه ای بین reviews\_per\_month و neighbourhood\_group هست ؟





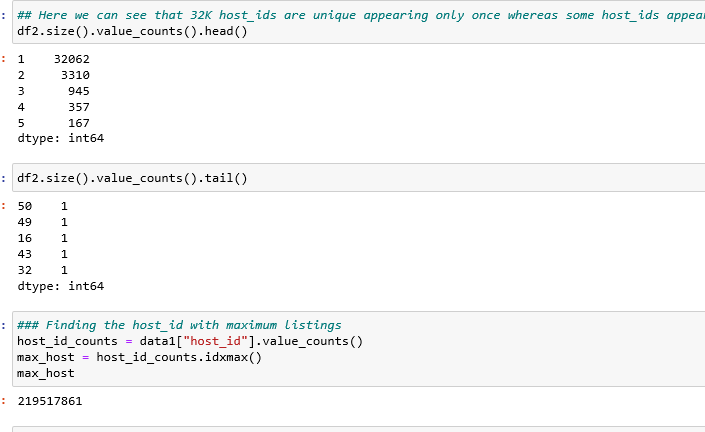
ما تا اینجا متوجه شدیم که manhatan و brooklin بالاترین لیست با قیمت بالای 150 دلار را دارا هستند و همین طور قیمت های بالای 100 دلار کل خانه ، اتاق خصوصی هست و کمترین مقدار اتاق مشترک است

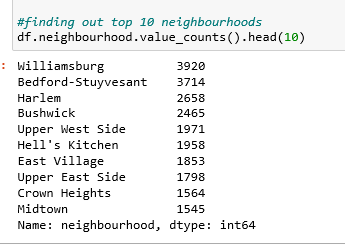
حالا در اینجا می خواهیم ماکسیمم host \_id را پیدا کنیم .

df2 = data1.groupby(["host\_id"])

max(df2.size())

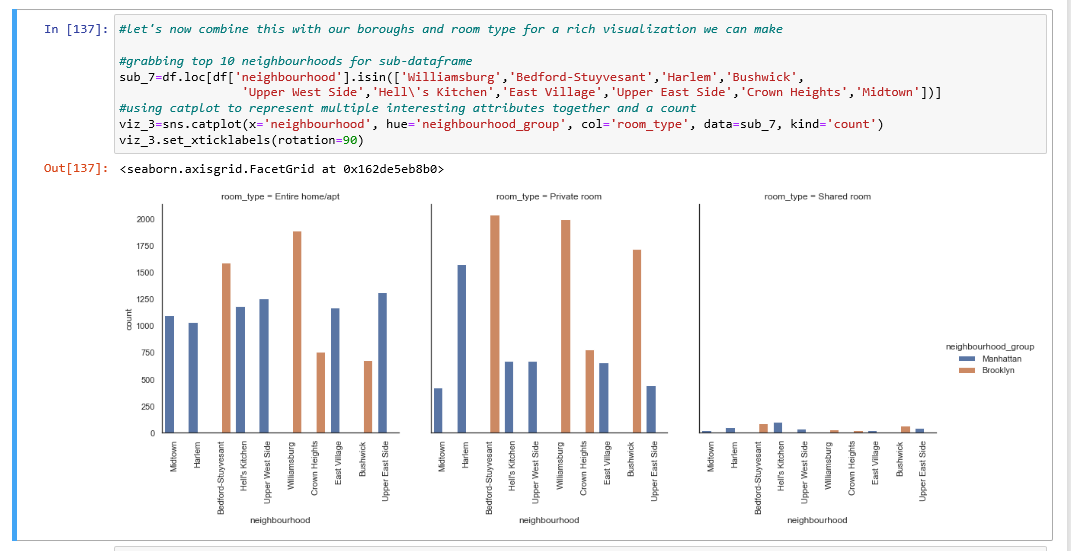
حالا پنج عنصر ابتدایی و انتهایی را فرا می خوانیم



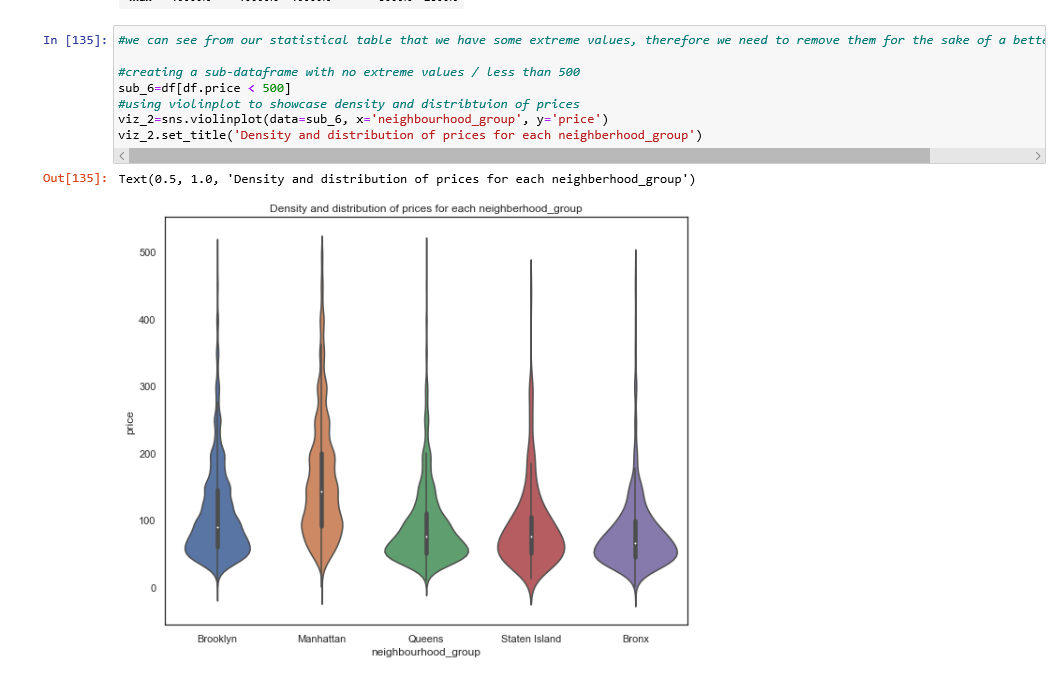
10 محله برتر که بیشترین لیست را دارند کدامند ؟

• Is there any noticeable difference of traffic among different areas and what could be the

reason for it?



مشاهداتی که قطعاً بیشترین تضاد را دارند این است که لیست Airbnb از نوع "Shared room" به سختی در میان 10 محله پرجمعیت موجود است. سپس ، می توانیم ببینیم که برای این 10 محله فقط 2 بخش ارائه شده است: منهتن و بروکلین. تا حدودی انتظار می رفت چون منهتن و بروکلین یکی از پر مسافرت ترین مقصد هستند ، بنابراین بیشترین لیست را در دسترس خواهند داشت. همچنین می توانیم مشاهده کنیم که بدفورد-استویوزانت و ویلیامزبورگ محبوب ترین شهرها برای بخش منهتن و هارلم برای بروکلین هستند.



با داشتن یک جدول آماری و یک طرح ویولن ، ما قطعاً می توانیم چند مورد را در مورد توزیع قیمت Airbnb در بخش های NYC مشاهده کنیم. اول ، می توانیم بگوییم که منهتن بالاترین طیف قیمت را برای لیست ها با قیمت 150 دلار به عنوان مشاهده متوسط دارد و پس از آن بروکلین با 90 دلار برای هر شب. کوینز و استاتن آیلند توزیع های بسیار مشابهی دارند ، برونکس ارزانترین از همه است. این توزیع و تراکم قیمت ها کاملاً انتظار می رفت. به عنوان مثال ، هیچ کس راز ندارد که منهتن یکی از گرانترین مکانهای زندگی در جهان است ، از طرف دیگر به نظر می رسد که برانکس دارای استاندارد زندگی پایین تر است

آیا رابطه ای میان host و مناطق پنج گانه هست ؟

• What can we learn about different hosts and areas?

