محمد عبدالهی

400422133

تمرین اول:

در این تمرین ما به بررسی و تحلیل دیتا New York City Airbnb Open Data میپردازیم. این دیتاست شامل اطلاعات مربوط به خانه های برای اجاره در شهر نیویورک بوده و پس از انجام مراحل پاکسازی داده سعی در تخمین قیمت خانه براساس فیچرهای موجود داریم.

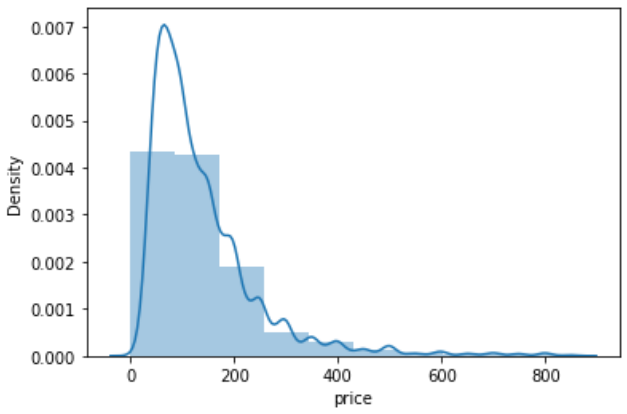
تسک های اصلی:

1. در بخش اول به کارهایی همچون لود کردن دیتا و پاکسازی دیتا و پر کردن مقادیر خالی و حذف فیچرهای بدون اطلاعات و حذف داده های پرت پرداختیم. کدها و پیاده سازی ها به طور کامل در نوت بوک پیاده سازی شده و شامل کامنت برای توضیحات بیشتر میباشد.

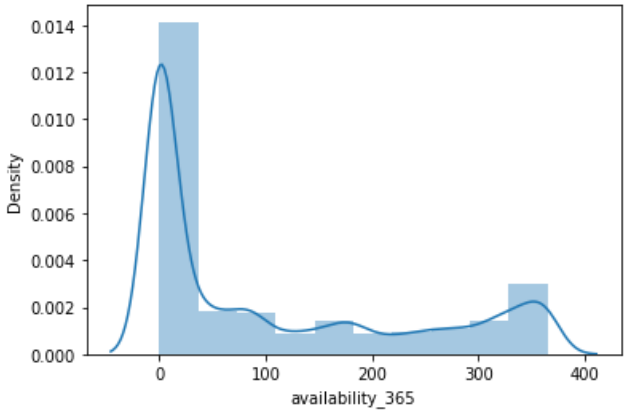
2. Data visualization: در این بخش سعی داریم تا با مصورسازی داده اطلاعات کلی درباره فیچرها و فهم بهتر آنها داشته باشیم.

   در گام اول چند مورد از فیچرهای continuous را نمایش داده و توزیع آنها را بررسی میکنیم:

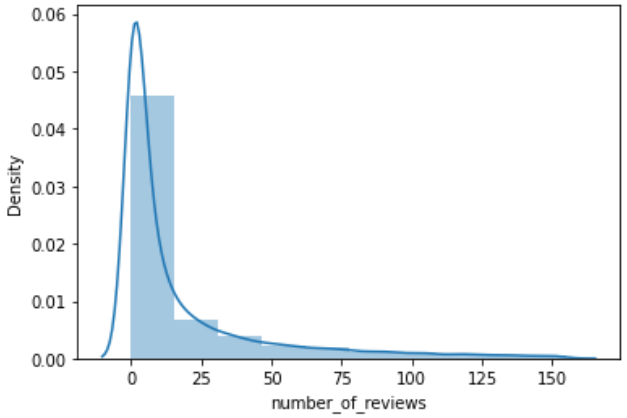
   price:



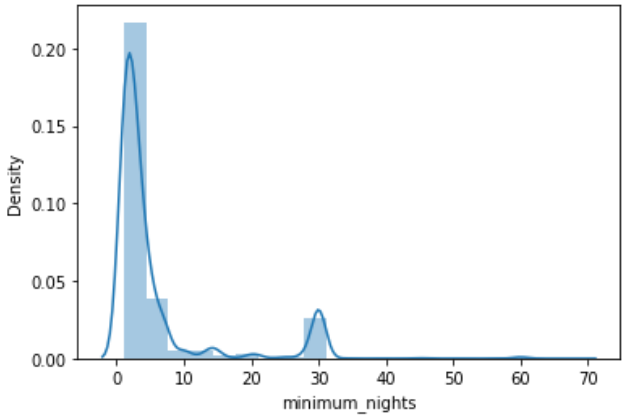
Availability_365:



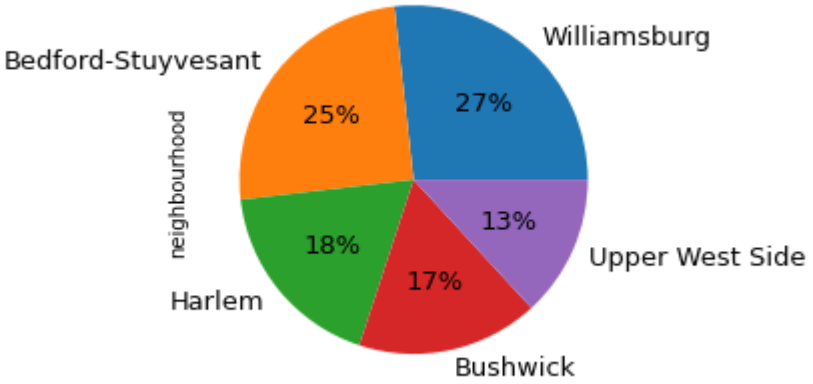
Number_of_reviews:



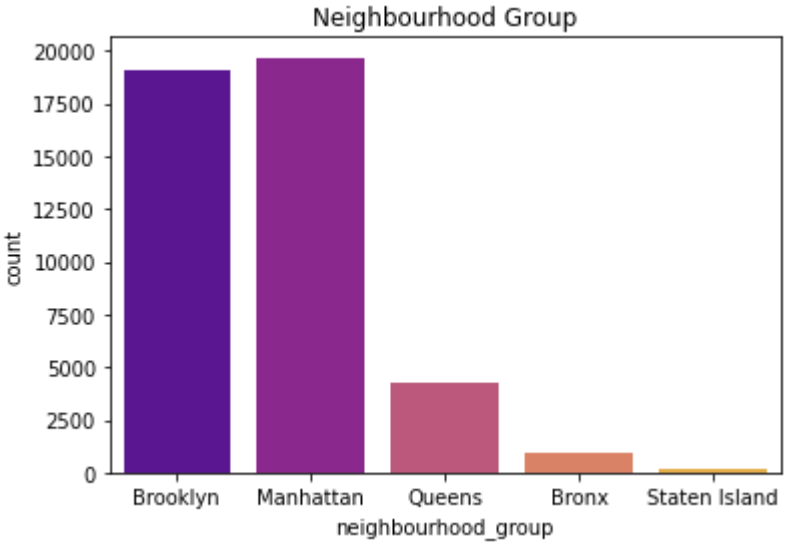
:Minimum_night



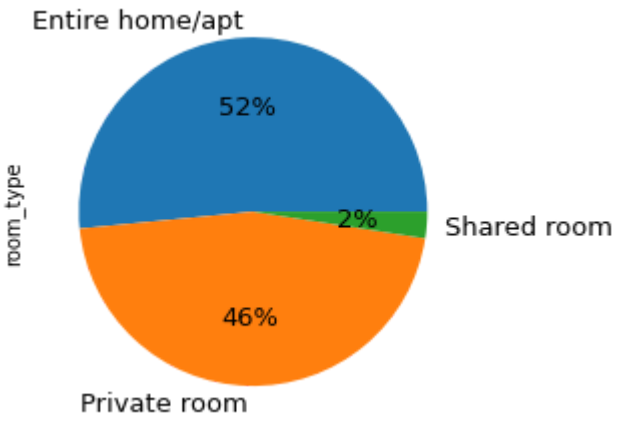
در مرحله بعد پای چارت مربوط پنج محله با بیشترین تعداد اگهی را نمایش می دهیم:



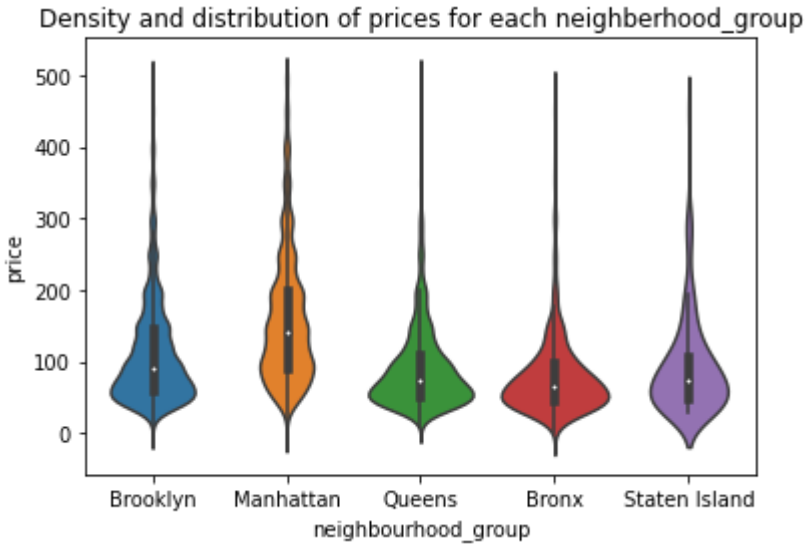
در نمودار بعد نمودار میله ای مربوط به فیچر neighbourhood group را رسم کرده که توزیع داده در محله های مختلف را نشان میدهد:



پای چارت بعدی مربوط به توزیع فیچر room type بوده که نوع خانه ها را نمایش میدهد

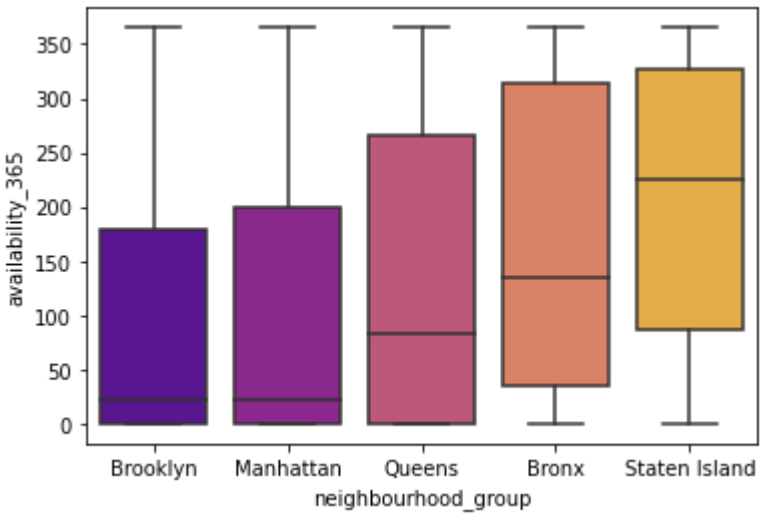


در violin plot شکل بعد سعی داریم توزیع قیمت در محله های مختلف را بررسی کنیم



همانطور که در شکل مشاهده میکنید manhattan دارای لاترین رنج قیمت بوده و میانگین قیمت خانه نیز در این محله بالاتر از بقیه محله ها می باشد پس از این محله هم brooklyn می باشد. همچنین queens و bronx قیمت ها تشابه بسیار زیادی دارند.

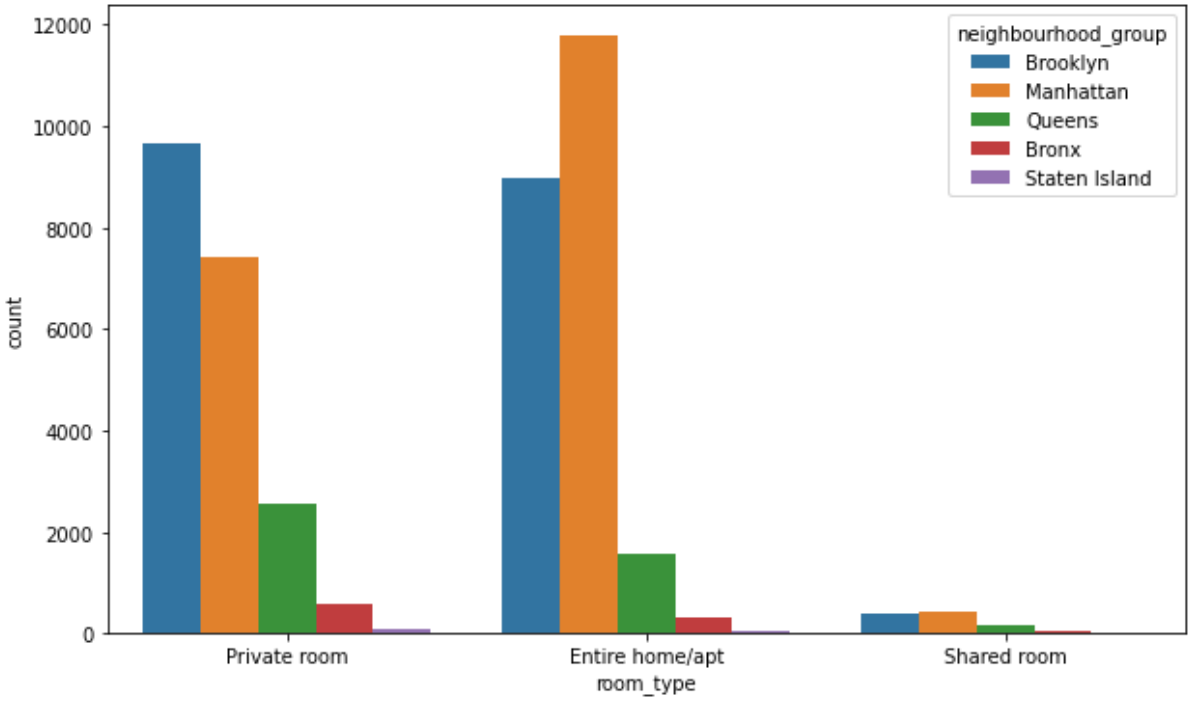
در شکل بعد نمودار جعبه ای مربوط به availability 365 و neighbourhood group آورده شده



همانطور که در تصویر پیداست خانه های staten island بیشتر از بقیه ها در دسترس بوده

و رنج در دسترس بودن بالاتری دارند این مقدار در brooklyn و manhattan کمترین میانگین را داراست.

در نمودار زیر مشاهده می کنیم که توزیع نوع خانه ها بر اساس محله چگونه است که بیشتر اگهی های private room در brooklyn بوده و بیشتر entire home ها در manhattan.



در نمودارهای بعد هم سعی شده تا توزیع فیچرهای مختلف بر روی نقشه و بر اساس latitude و longitude نمایش داده شده که میتوانید در کد مشاهده بفرمایید.

3. در سوال سه خواسته شده که میزبان ها با بیشترین تعداد آگهی یافته شود که برای این منظور تابع `get_hosts_with_most_house` را پیاده سازی کردیم که دیتاست اصلی را خوانده و نتیجه خواسته شده را محاسبه میکند.

4. در سوال بعدی از ما خواسته شده تا با در نظر گرفتن number of reviews به عنوان یک معیار رضایت صاحبان آگهی با بیشترین رضایت یافته شود که برای این منظور تابع را پیاده سازی کردیم.

پس از بررسی نتایج بدست آمده میتوان نتیجه گرفت که این صاحبان آگهی خانه هایی با قیمت کمتر نسبت به بقیه خانه ها در محله های خود را دارا هستند که به صورت private room میباشند.

5. در بخش از ما خواسته شده که پنج تست فرض دلخواه را مطرح و نتایج آنها را تحلیل کنیم. توضیحات و کدهای این بخش را در نوت بوک میتوانید مشاهده بفرمایید.

6. در این بخش با استفاده از دیتاست پاکسازی شده سعی در مدل سازی و تخمین قیمت براساس این فیچرها داشتیم. لازم به ذکر است که توزیع لیبل price دارای مقداری چولگی به راست بود و برای حل این موضوع مقدار لگاریتم price را به عنوان لیبل در نظر گرفتیم(تصویر توزیع قبل و بعد از عملیات در کد ضمیمه شده).

در حالت اول فیچرهای کتگوریکال را به صورت one hot انکود کرده و با الگوریتم های linear regression, lasso regression, ridge regression, gradient boosting سعی در تخمین قیمت داشتیم.

در مرحله دوم با استفاده از pca و select k best features سعی در مهندسی فیچرها و بهبود عملکرد مدل داشتیم و توانستیم ارور مدل را کاهش داده و تخمین های بهتری داشته باشیم.

در نهایت بهترین نتیجه بدست آمده به صورت زیر بود:

```
Mean absolute error : 0.301022
Root mean squared error : 0.405172
R2_score : 0.613936
```

7. تسک امتیازی شماره دو:

بررسی تاثیر زن یا مرد بودن در قیمت و رضایت

برای این منظور از یک کتابخانه gender-guesser برای تشخیص زن یا مرد بودن صاحبان آگهی ها استفاده کرده ایم و از روی اسم کوچک آنها این مورد را تشخیص داده ایم.
سپس بر اساس جنسیت به دست و مصورسازی و تست فرض بررسی کرده که آیا زن یا مرد بودن صاحب اگهی نقشی در رضایت یا قیمت داشته است. و در نهایت به این نتیجه میرسیم که جنسیت صاحب اگهی تاثیری در قیمت یا رضایت مشتریان نداشته است.