به نام خدا

گزارش تمرین های سری اول درس داده کاوی

استادان گرامی:

جناب آقای دکتر فرهانی و جناب آقای دکتر خردپیشه

دستیار آموزشی : جناب آقای شریفی

گردآورنده: سلاله شیخیان

مقدمه

در این فایل به تحلیل دو مجموعه داده با استفاده از استنتاج آماری می پردازیم. به سوالات مطرح شده زیر پاسخ می دهیم. و بر اساس داده ها سوالات دیگری می پرسیم و پاسخ آن ها را می یابیم.

مجموعه داده اول:

کد های این بخش در فایل Project1_ab_nyc.ipynb قرار دارد

خانه های اجاره داده شده برای اقامت کوتاه مدت در شهر نیویورک آمریکا. تعداد کل: ۴۸۸۹۵

What can we learn about different hosts and areas?

برای بدست آوردن اطلاعات از روی منطقه ها ابتدایی ترین کار تحلیل اطلاعات اولیه و رسم نمودار های آن هاست:

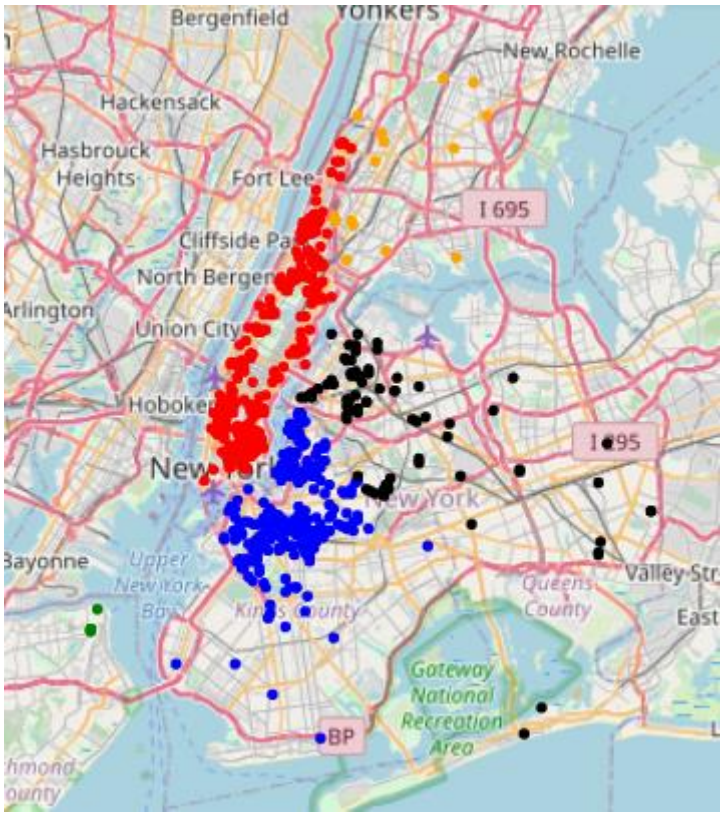
تعداد داده ها (row) در هر ناحیه

manhattan	21661	قرمز
brooklyn	20104	آبی
queens	5666	سیاه
bronx	1091	زرد
Staten island	373	سبز

همانطور که مشاهده می کنیم تراکم اجاره ها در ناحیه ها با هم تفات زیادی دارد و بیشتر خانه ها در دو ناحیه اول اجاره داده شده اند.

مشاهده داده ها در نقشه

از آنجایی که تعداد داده ها برای نمایش به صورت داینامیک روی نقشه زیاد است، تعدادی از داده ها را به گونه ای انتخاب می کنیم که نسبت داده های هر ناحیه نیز، در نقشه نمایش داده شود. پس داده ها را بر اساس ناحیه دسته بندی می کنیم و از هر دسته به صورت تصادفی یک هفتادم داده ها را انتخاب می نماییم. و بر اساس طول و عرض جغرافیایی روی نقشه نمایش می دهیم. هر شهر با یک رنگ نمایش داده شده است تا بهتر بتوان تفاوت پراکندگی و توزیع خانه های اجاره ای در هر منطقه را مشاهده کرد.



علاوه برنحوه پراکندگی خانه ها بین مناطق مختلف، می توانیم پراکندگی خانه ها داخل هر منطقه را مشاهده کنیم. در منطقه آبی رنگ یعنی Brooklyn در قسمت شمالی منطقه تراکم بیشتر است اما در منطقه زرد رنگ یعنی bronx تراکم تقریبا یکنواخت است.

بخش دوم سوال:

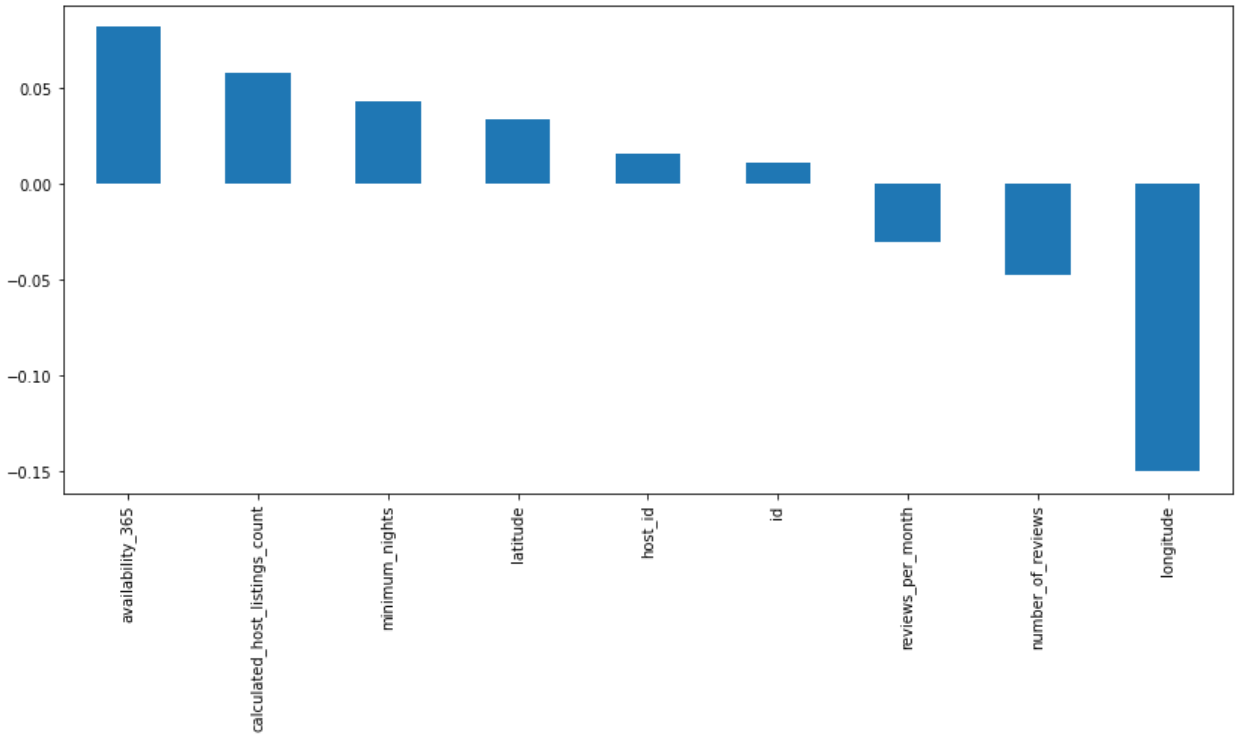
چه اطلاعاتی از طریق اجاره دهندگان بدست می آید؟

ابتدا محاسبه می کنیم که هر میزبان چند خانه را برای اجاره گذاشته است. برای این کار ابتدا حالت های تکراری host-id را از حذف می کنیم و تعداد باقی مانده را بدست می آوریم. این تعداد 37457 است که تعداد اجاره دهندگان را نشان می دهد. از آنجایی که این تعداد بسیار زیاد است و تقریبا سه چهارم از داده ها را تشکیل می دهد، نتیجه می گیریم که بیشتر اجاره دهندگان تنها یک بار از طریق سایت ما خانه خود را برای اجاره گذاشته اند.

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

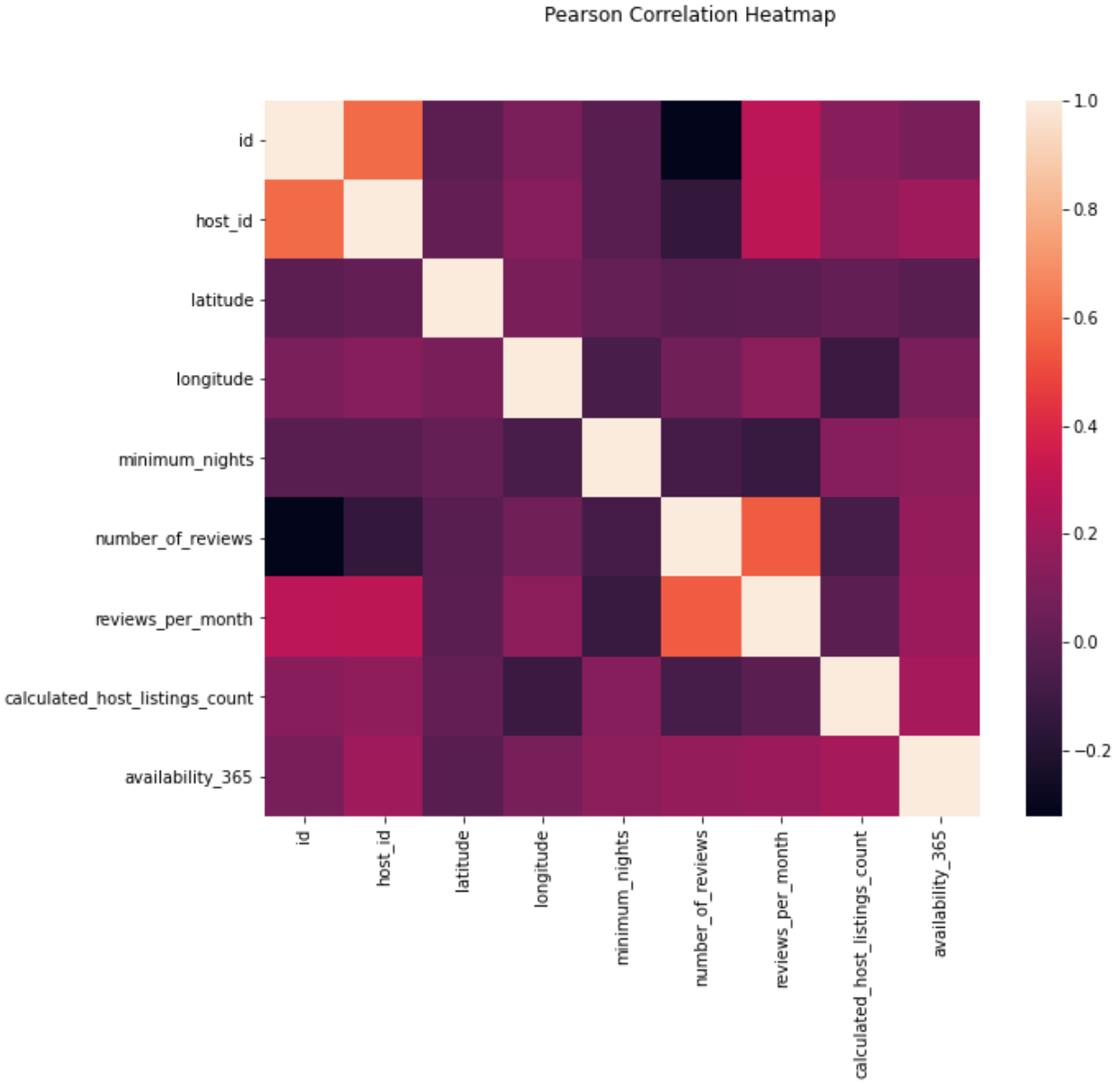
پیشبینی قیمت :

برای این کار میخواهیم از رگرسیون خطی استفاده کنیم.
ابتدا همبستگی قیمت با سایر نمودار ها را حساب می کنیم.



در این نمودار بیشترین همبستگی با مختصاف وجود دارد که این مسئله منطقیست زیرا میدانیم محله روی قیمت تاثیرگذار است. بعد از آن تعداد بازدید و calculated_host_listings_countبیشترین همسبتگی را دارند.

همچنین از نمودار حرارتی هم می توان میزان همبستگی را مشاهده کرد

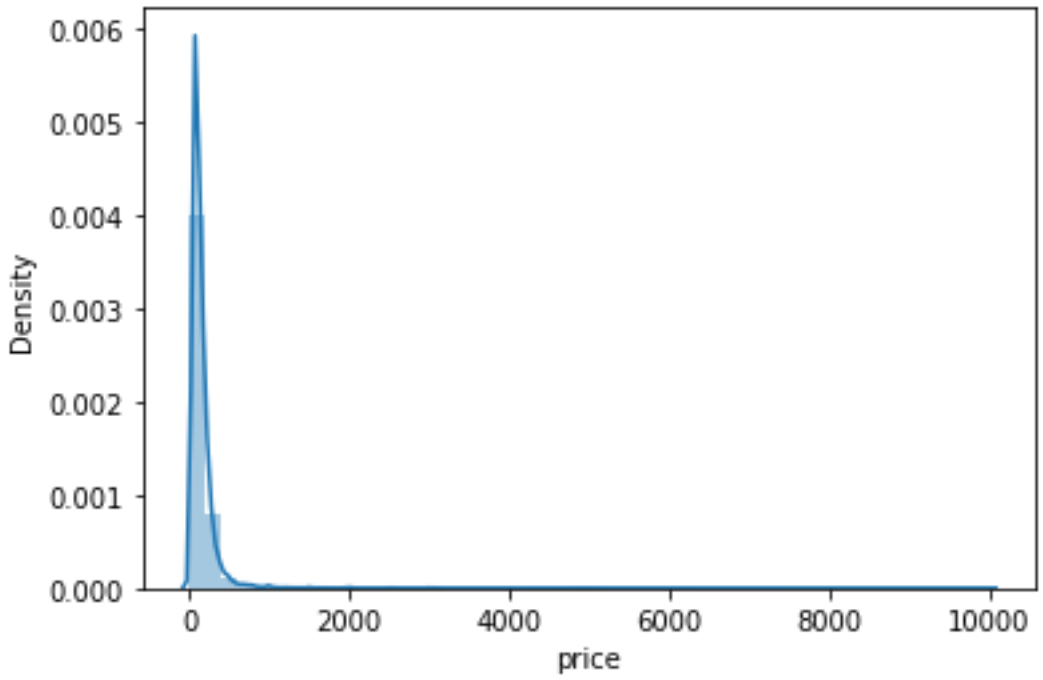


.

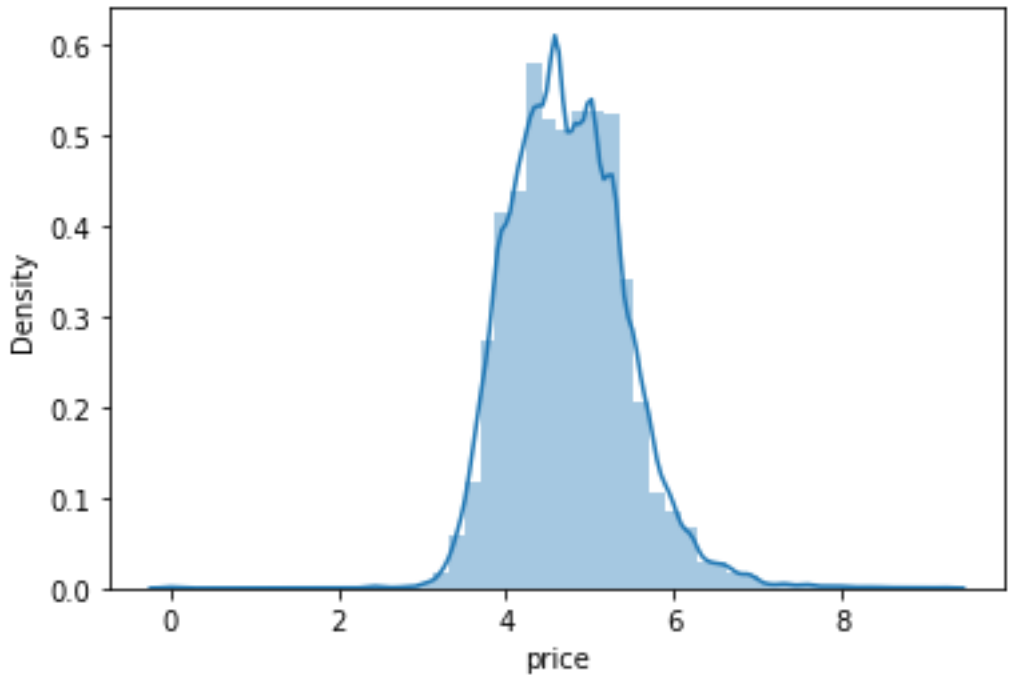
دو ویژگی تعداد بازدید در ماه و تعداد بازدید همبستگی زیاد دارند زیرا درواقع هردو به یک موضوع اشاره می کنند. پس در صورت استفده باید تنها از یکی از آنها استفاده نماییم. البته الگوریتم های برای تشخیص و بررسی این حدس ها وجود دارد که میتوان از آنها هم استفاده نمود.

با استفاده از این نمودار ها چهار متغیر گفته شده را انتخاب می نماییم.

حال برای قیمت ابتدا نمودار توزیع آن را رسم می کنیم:



از آنجایی که طبق نمودار های رسم شده توزیع، توزیع قیمت نرمال نیست آن را نرمال می کنیم زیرا . از آنجا که مدل‌های خطی با داده‌های دارای توزیع نرمال کار می‌کنند، در اینجا Price تبدیل و توزیع آن نرمال‌تر می‌شود



داده های آموزش و تست را هم جدا میکنیم. رگرسیون خطی را با استفاده از
کتابخانه `sklearn` با داده های آموزشی آموزش میدهیم و درنهایت قدرمطلق میزان خطا را
بدست می آوریم .
`8.670310030717 22e-14`

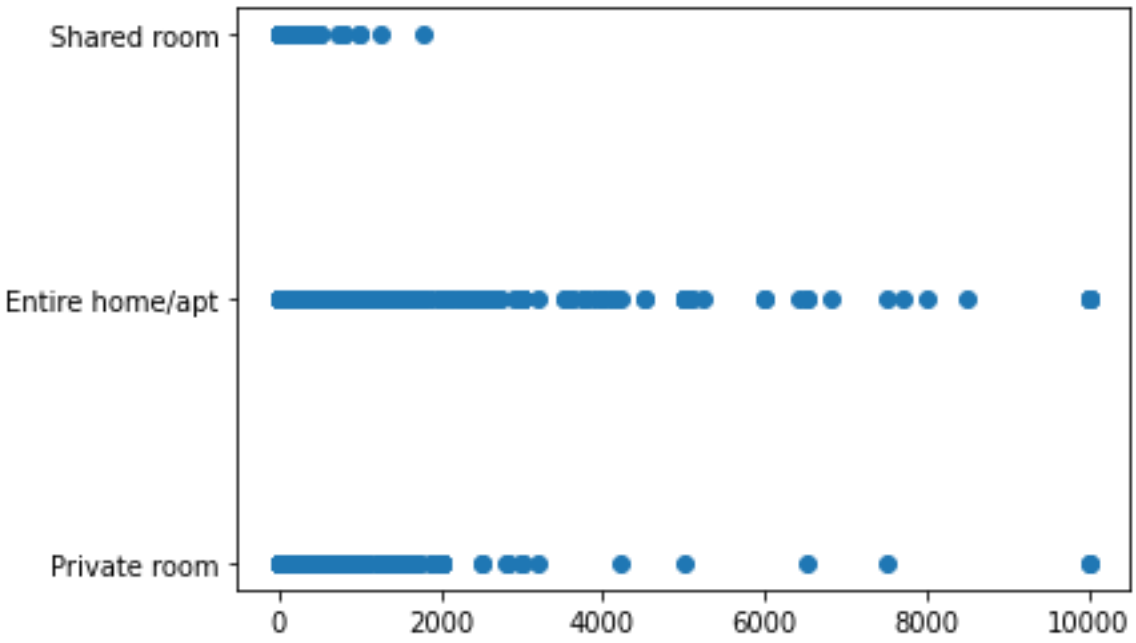
این بخش از کد را به صورت جداگانه در یک فایل `priceRegression` .قرار داده ام.

- *Which hosts are the busiest and why?*
- *Is there any noticeable difference of traffic among different areas and what could be the reason for it?*

برای بدست آوردن ترافیک در نواحی مختلف، می توانیم از تعداد بازدید های ماهانه و
calculated_host_listings_count استفاده کنیم .

نمودارها (مصور سازی داده ها)

نمودار قیمت بر اساس نوع اتاق

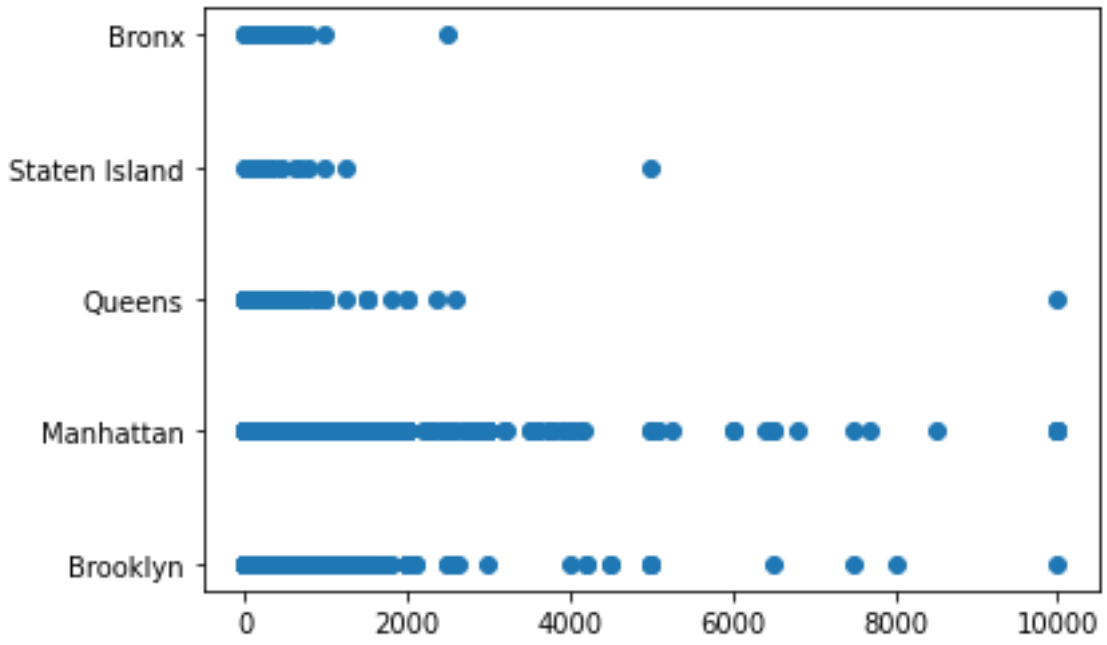


با توجه به این نمودار میتوان بیشینه قیمت برای هر نوع خانه اجاره ای را مشاهده کرد. همچنین مشخص است که در هر دسته بیشتر داده ها در کدام بازه قرار دارند و به نحوی میتوان پراکندگی قیمت را هم مشاهده کرد اما توزیع را نشان نمیدهد زیرا تعداد دقیق در هر قیمت را نشان نمیدهد. برای مشاهده دقیق توزیع قیمت از نمودار های دیگری استفاده خواهیم نمود.

میانگین قیمت هر نوع خانه

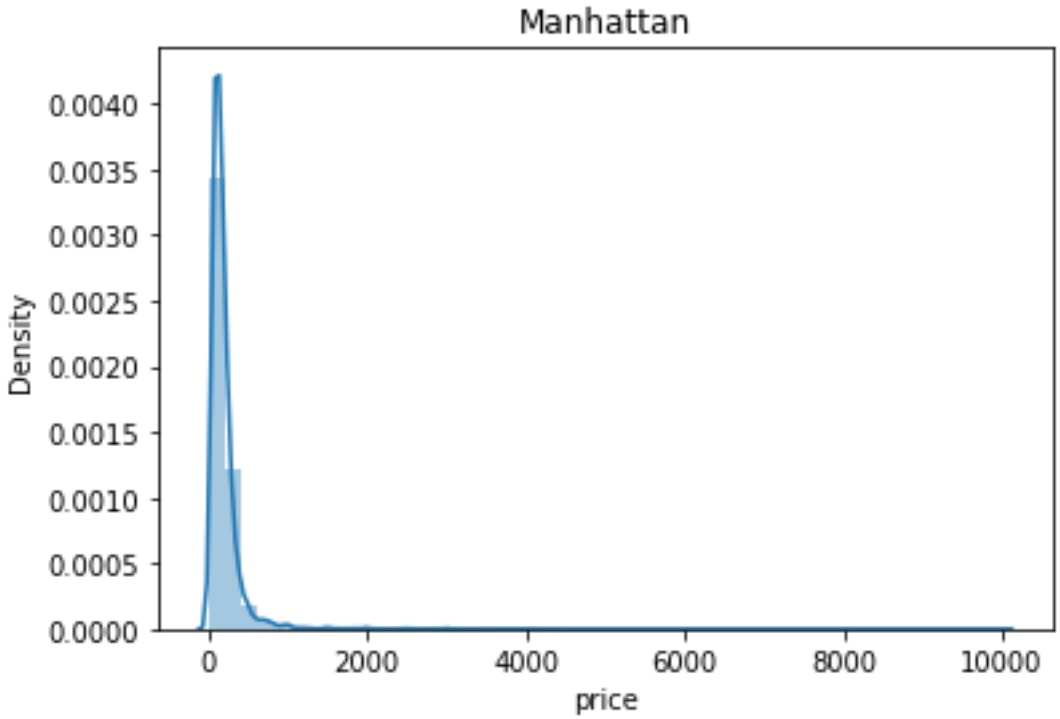
توزیع قیمت در هر نوع اتاق

نمودار قیمت بر اساس منطقه



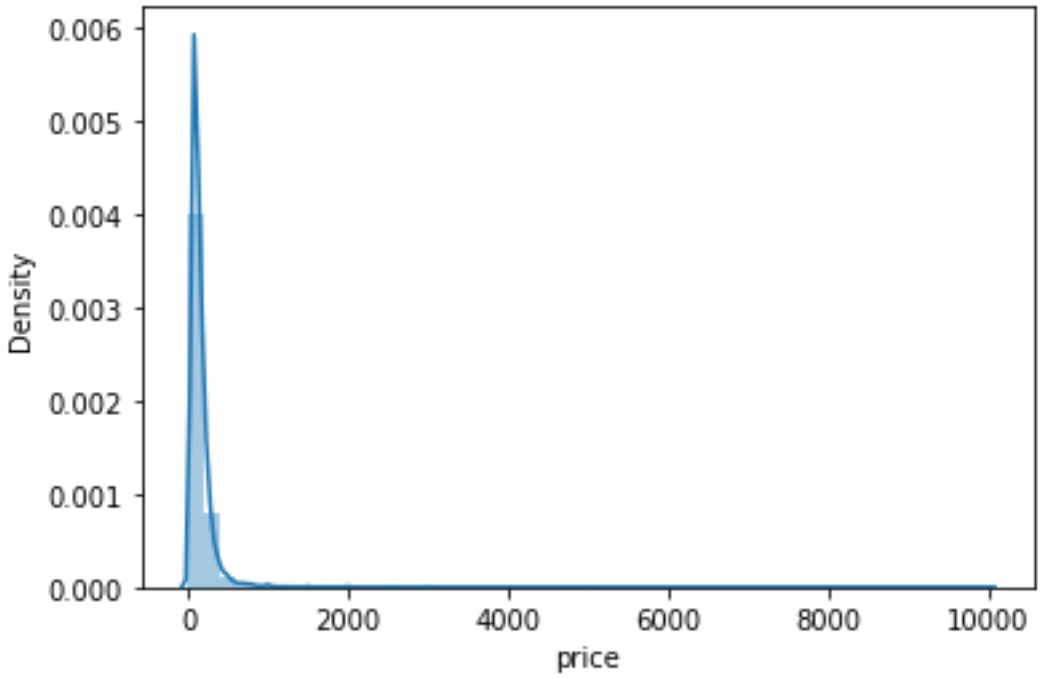
این نمودار هم نمیتواند تعداد را در هر قیمت نشان دهد پس با استفاده از نمودار توزیع زیر برای هر منطقه نمودار را رسم میکنیم.

اگر بخواهیم نمودار ها را بدون محدود کردن دامنه نمایش دهیم به شکل زیر در می آیند:

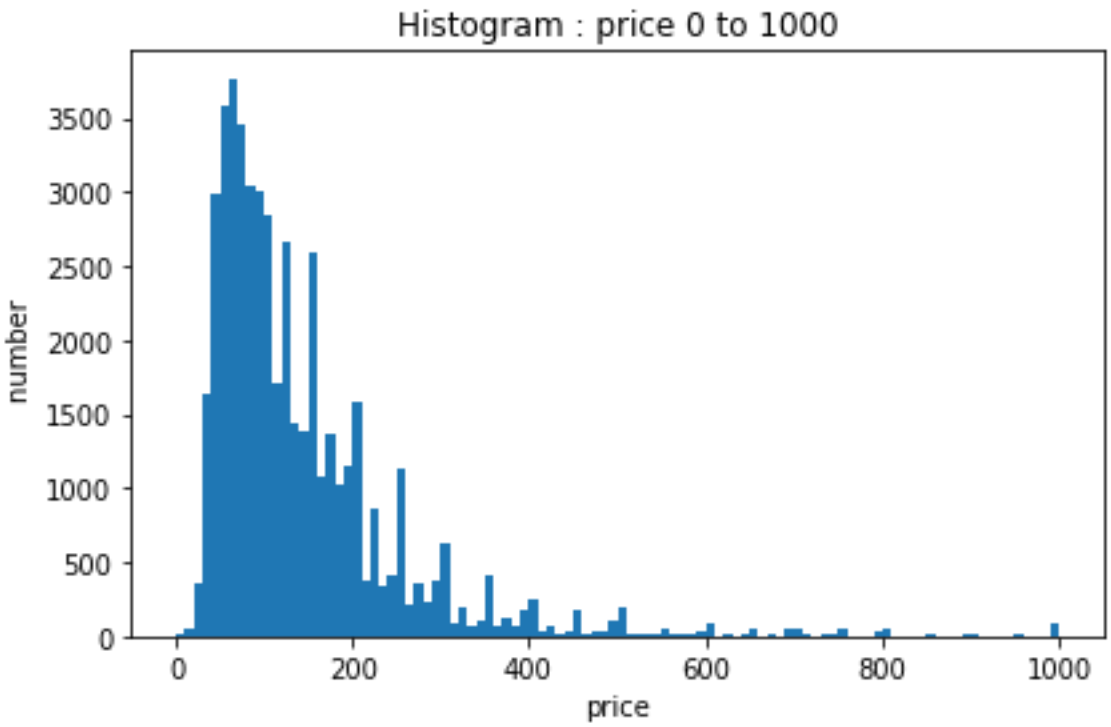


در نتیجه از داده های با قیمت بالا که تعدادشان هم کم هست صرف نظر میکنیم تا بهتر بتوانید نمودار توزیع را ببینیم.

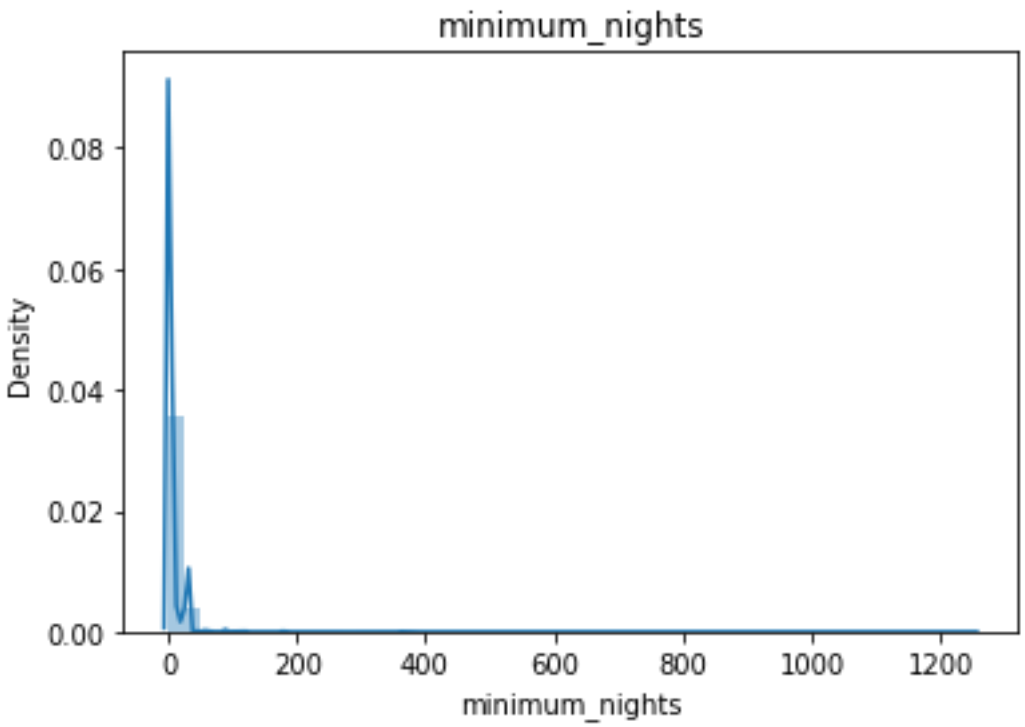
نمودار توزیع قیمت کل



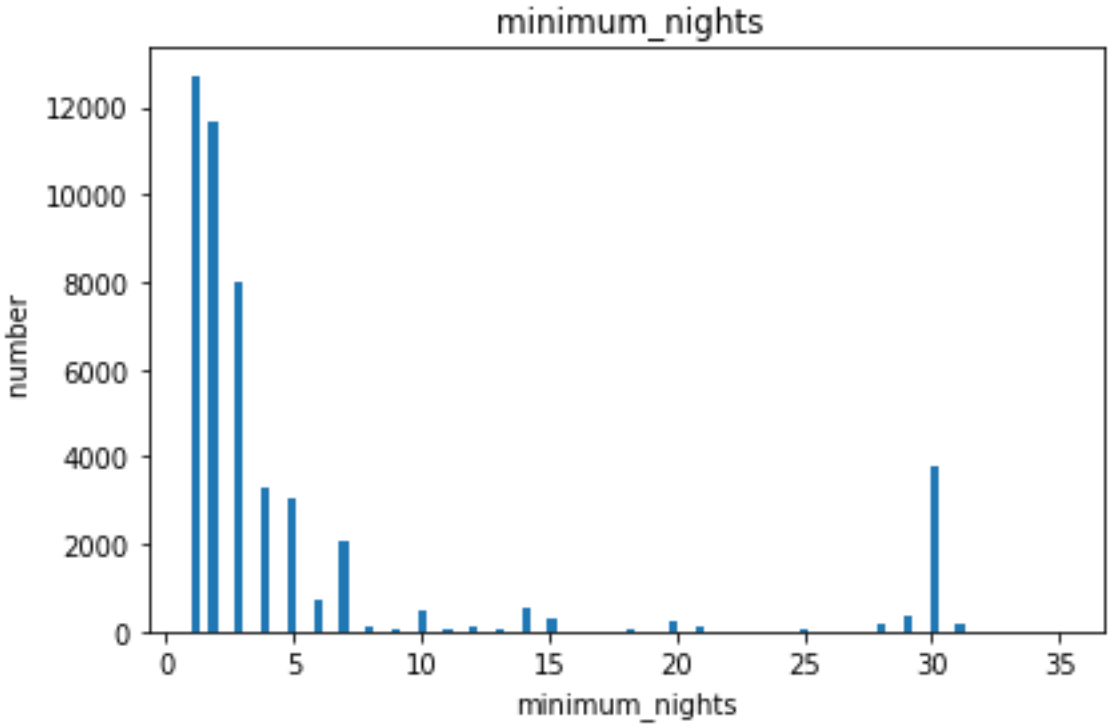
همانطور که از این نمودار توزیع پیداست بیشتر داده ها در بازه صفر تا هزار قرار دارند. پس برای واضح تر شدن نمودار بازه را به این قسمت محدود می کنیم و از نمودار هیستوگرام استفاده می نماییم:



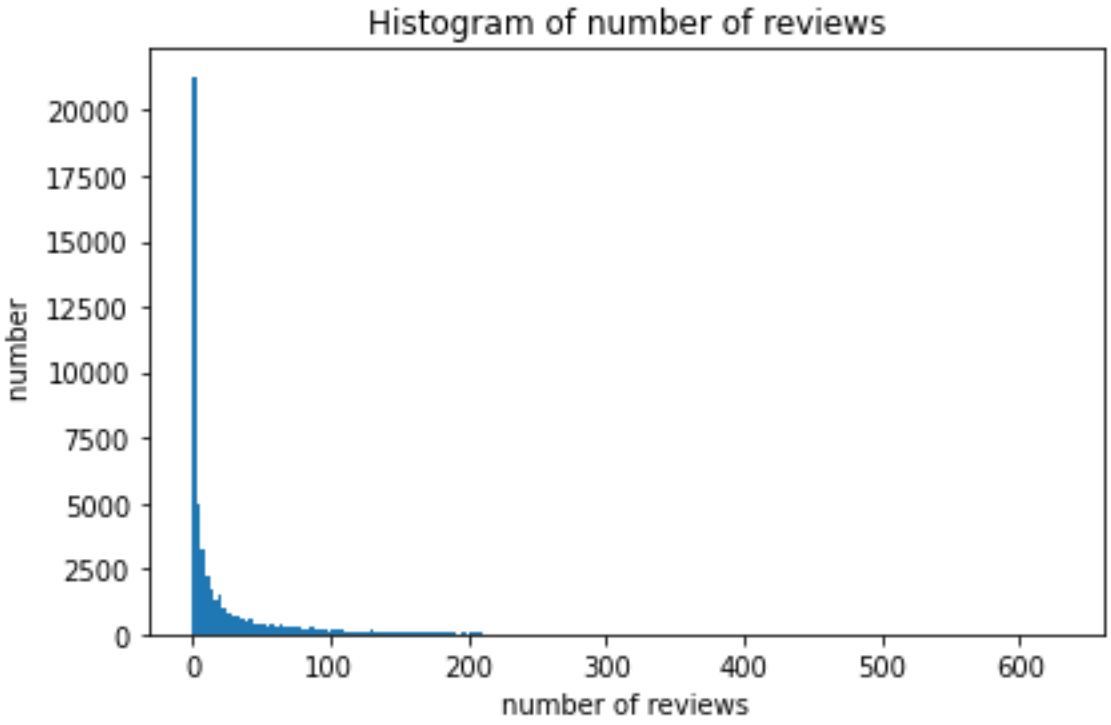
نمودار توزیع مینیمم شب ها



یا به شکل دقیق تر:

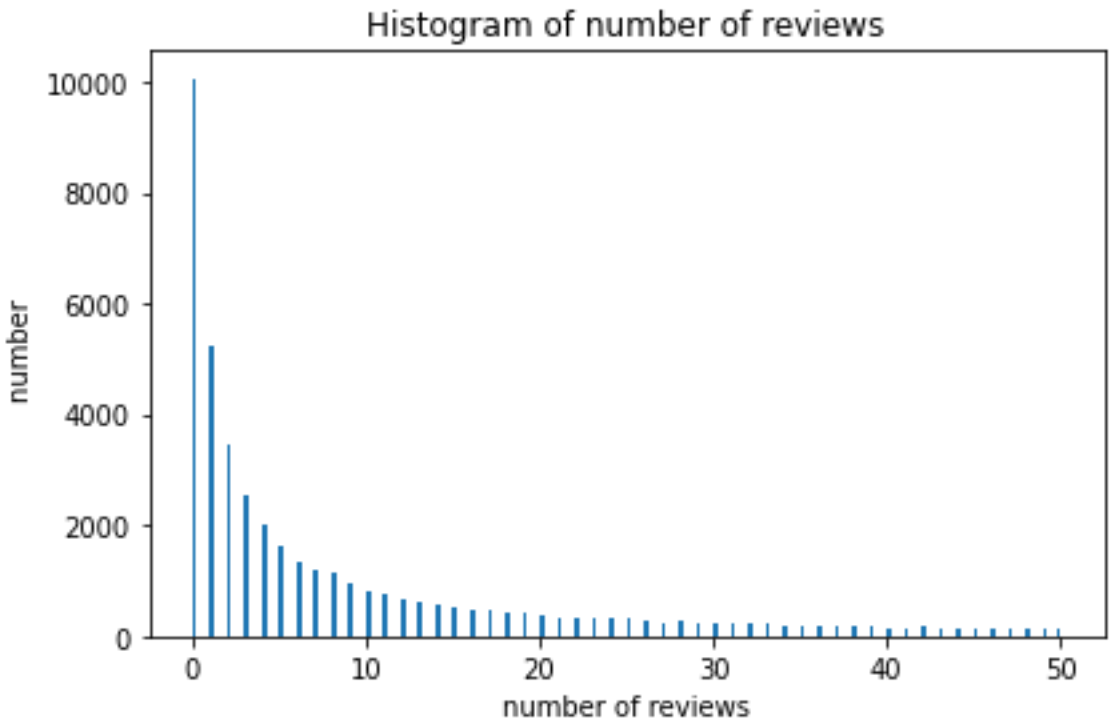


نمودار توزیع تعداد بازدید ها در حالت کلی

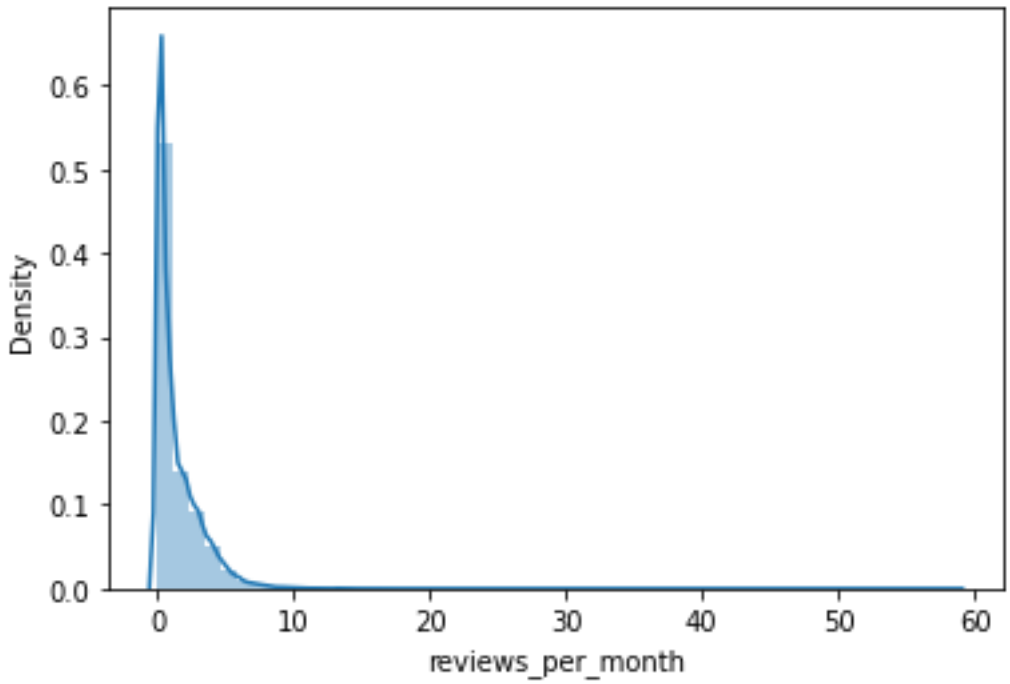


نمودار توزیع تعداد بازدید در بازه صفر تا ۵۰ بازدید

این نمودار نشام میدهد که حدود یک پنجم خانه هایی که برای اجاره گذاشته شده اند هیچ بازدیدی نداشته اند.



تعداد بازدید ها در ماه



سوال: چرا حدود یک پنجم خانه ها بازدیدی نداشته اند؟

از آن جایی که نمودار توزیع تعداد بازدید دارای شکل مشخصی است که هر چه عدد بازدید بزرگتر باشد تعداد کمتر میشود، به نظر این تعداد زیاد برای بازدید صفر تا حدودی طبیعی بنظر می آید.

اما چون اختلاف آن با تعداد بازدید یک بسبار زیاد است بهتر است عوامل دیگر را هم بررسی نماییم.

به نظر میرسد که عوامل زیادی میتواند بر بازدید نداشتن یک خانه تاثیر داشته باشد . مانند قیمت بالا، منطقه قرار گرفتن خانه، مینیمم شب های قایل اجاره ی بالا و ... .

در این قسمت برخی از این عوامل را بررسی میکنیم که آیا بر اساس نمودار ها میتوانیم رابطه ای بین تعداد بازدید و قیمت و یا رابطه بازدید و محل قرار گیری خانه ها و یا رابطه تعداد بازدید و مینیمم شب ها پیدا کنیم یا خیر. و اگر ارتباطی وجود داشت بتوانیم بعد تر با آزمون های فرض درستی حدسمان را بررسی کنیم.

بررسی این که آیا رابطه ای بین قیمت و تعداد بازدید وجود دارد یا خیر:

ابتدا سعی می کنیم داده ها را در نمودار قرار دهیم تا ببینیم آیا رابطه محسوسی بین قیمت و تعداد بازدید دیده می شود یا خیر.

داده ها را بر اساس تعداد بازدید به سه دسته تقسیم می کنیم:

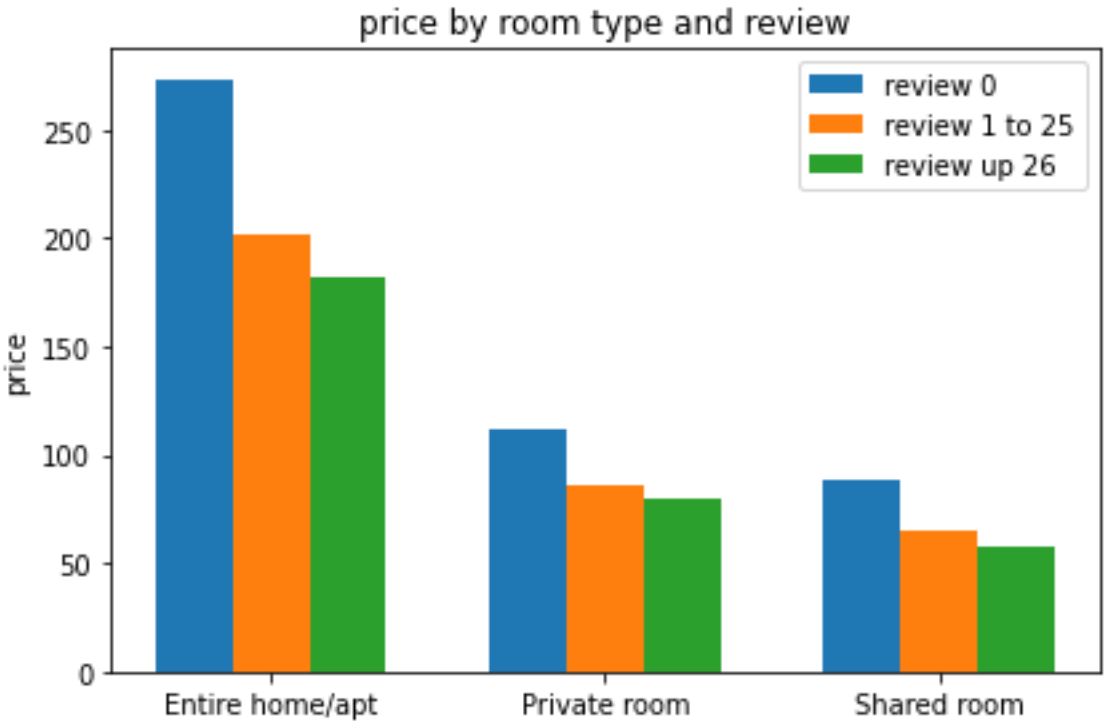
- دسته اول: خانه های بدون بازدید
- دسته دوم : خانه های با بازدید بین ۱ تا ۲۵
- دسته: خانه های با بازدید بیش از ۲۶
- از طرفی خانه ها دارای ۳ نوع مختلف هستند که عبارت اند از:

- `Entire home/apt`

- Private room
- Shared room

میخواهیم ببینیم که میانگین قیمت در هر دسته تعداد بازدید، برای هرکدام از این سه نوع خانه اجاره ای چقدر است.

پس لازم است هر سه دسته بازدیدمان را بر اساس نوع خانه گروه بندی کنیم و میانگین قیمت هر گروه را بدست آوریم. سپس با استفاده از نمودار میانگین قیمت را نمایش دهیم تا ببینیم آیا رابطه به وضوح روشنی بین قیمت ها هست ؟



نمودار بالا رابطه ای بین قیمت و تعداد بازدید را در هر سته نوع خانه نشان میدهد.

به طوری که هر نوع خانه ای را که در نظر بگیریم(مثلا دسته entire home) خانه هایی که بازدید صفر داشته اند، بیشترین قیمت را نیز داشته اند. و از طرفی خانه هایی که بازدید بیش از ۲۶ داشته اند، به نسبت دو بازه دیگر میانگین قیمت کمتری دارند. اما این هنوز یک حدس است و میتوان این رابطه را با آزمون های مختلف آزمود.

سوال: چرا در نمودار هیستوگرام قیمت، قیمت صفر داریم؟ تعداد آن ها چقدر است و دیگر ستون های آن چه مقادیری دارند؟

تعداد خانه های با قیمت صفر ۱۱ عدد است. سایر اطلاعات این سطر ها طبیعی است و به نظر می رسد که این داده ها دارای خطا شده اند و قیمت صفر ثبت شده است.

مجموعه داده دوم

تمام کد های این بخش در فایل project1-football قرار دارد

Who is the best team of all time

برای بدست آوردن بهترین تیم در همه زمان ها در یک حلقه فور برای هر برد به برنده امتیاز ۳ برای مساوی به هر یک امتیاز ۱ و برای باخت صفر امتیاز نسبت دادیم و برای هر تیم امتیازات را جمع کردیم. طبق نتیجه بهترین تیم در همه ی زمان ها برزیل بدست آمد.

Which teams dominated different eras of football

دوره های مختلف فوتبال به چند دوره زیر تقسیم میشوند:

The Golden Age	۱۸۷۰تا شروع جنگ جهانی اول
Emerge of continental era	۱۹۱۶تا ۱۹۳۰
Inter-war Football in Europe	۱۹۳۰ تا ۱۹۳۷
Post WWII soccer era	۱۹۳۷ به بعد

داده ها را بر اساس این دوره ها جدا مینماییم و برای هر دوره دستور بالا را اجرا مینماییم.نتیجه:

England	۱۸۷۰تا شروع جنگ جهانی اول	**The Golden Age**
Argentina	۱۹۱۶تا ۱۹۳۰	**Emerge of continental era**
Germany	۱۹۳۰ تا ۱۹۳۷	**Inter-war Football in Europe**
Brazil	۱۹۳۷ به بعد	**Post WWII soccer era**

البته بهتر است به جای تکرار کد از تابع استفاده نماییم اما چون تعداد فقط ۴ مورد و کد کوتاه بود به صورت دستی هرکدام را حساب کردیم.

What trends have there been in international football throughout the ages – home-advantage, total goals scored, distribution of teams' strength etc

Which countries host the most matches where they themselves are not participating in

تیمی که بیشتر از باقی تیم ها میزبان بازی هایی بوده که در آن ها حضور نداشته تیم Mexico با میزبانی ۲۴۸ بازیست و بعد از آن برزیل بیشترین رتبه را دارد.

با یک دستور if برای تمامی ردیف ها به مقایسه کشور و دو تیم شرکت کننده پرداختیم. و هر بار که کشور میزبان با هر دو تیم مخالف بود یک واحد به شمارنده آن اضافه می کنیم و درنهایت بیشترین را بازمیگردانیم.

How much, if at all, does hosting a major tournament help a country's hances in the tournament

Which teams are the most active in playing friendlies and friendly tournaments - does it help or hurt them

برای این که بفهمیم کدام تیم بیشترین بازی دوستانه را داشته است، ابتدا بازی های دوستانه را جدا می کنیم سپس به ازای هر بازی دوستانه یک امتیاز به تعداد بازی های دوستانه هر تیم اضافه میکنیم تا بیشترین را بدست آوریم.

آلمان با ۵۷۲ بازی دوستانه بیشترین تعداد بازی دوستانه را داشته است.

حالا برای این که ارتباط تعداد بازی های دوستانه و برد های تیم در بازی های غیردوستانه را بدست آوریم باید داده ها را به بازه های یک ساله تقسیم کنیم. سپس تعداد بازی های دوستانه و تعداد برد ها را بدست آوریم.

تعداد بازی دوستانه و تعداد برد ها

نسبت زمان به تعداد برد

سوال: روند پیشرفت ۵ تیم برتر در طول دوره ای مختلف چگونه بوده است(بر حسب تعداد گل)

در هر سال چند بازی رابرده اند؟