**مقدمه**

میهمانان و میزبانان فراوانی از Airbnb برای گسترش امکان سفر استفاده می کنند. امروز ، Airbnb به یکی از خدماتی تبدیل شده که توسط تمام دنیا مورد استفاده قرار می گیرد. تجزیه و تحلیل داده ها در مورد میلیون ها لیست ارائه شده از طریق Airbnb یک عامل حیاتی برای این شرکت است. این میلیون ها لیست داده های زیادی تولید می کنند - داده هایی که می توانند برای امنیت ، تصمیمات تجاری ، درک رفتار و عملکرد مشتریان و ارائه دهندگان (میزبان) ، راهنمایی برای ابتکارات بازاریابی ، اجرای خدمات اضافی نوآورانه و موارد خیلی بیشتر در سیستم تجزیه و تحلیل و استفاده شوند

این مجموعه داده در حدود 49،000 مورد مشاهده با 16 ستون ویژگی در خود دارد و ترکیبی است از مقادیر ترکیبی و عددی.

***id int64***

***name object***

***host\_id int64***

***host\_name object***

***neighbourhood\_group object***

***neighbourhood object***

***latitude float64***

***longitude float64***

***room\_type object***

***price int64***

***minimum\_nights int64***

***number\_of\_reviews int64***

***last\_review object***

***reviews\_per\_month float64***

***calculated\_host\_listings\_count int64***

***availability\_365 int64***

***dtype: object***

پس از بارگذاری مجموعه داده می توان موارد مختلفی را مشاهده کرد. این 16 ستون مقدار بسیار غنی از اطلاعات را برای کاوش عمیق داده ها فراهم می کند. در حال حاضر برخی از مقادیر از دست رفته و nan را مشاهده میکنیم که نیاز به تمیز کردن دارند. بعداً ، ممکن است لازم باشد برای تجزیه و تحلیل و پیش بینی ، با نگاشت مقادیر خاص به یک و صفر ادامه دهیم.

تعداد مقادیر از دست رفته برای هر ستون و یا ویژگی:

***id 0***

***name 16***

***host\_id 0***

***host\_name 21***

***neighbourhood\_group 0***

***neighbourhood 0***

***latitude 0***

***longitude 0***

***room\_type 0***

***price 0***

***minimum\_nights 0***

***number\_of\_reviews 0***

***last\_review 10052***

***reviews\_per\_month 10052***

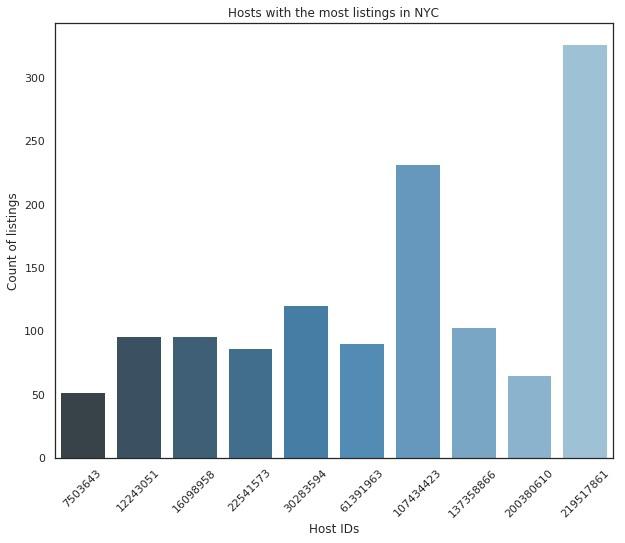
***calculated\_host\_listings\_count 0***

***availability\_365 0***

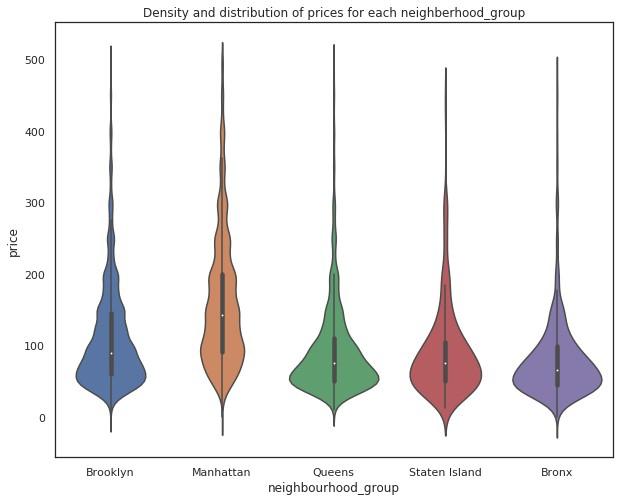
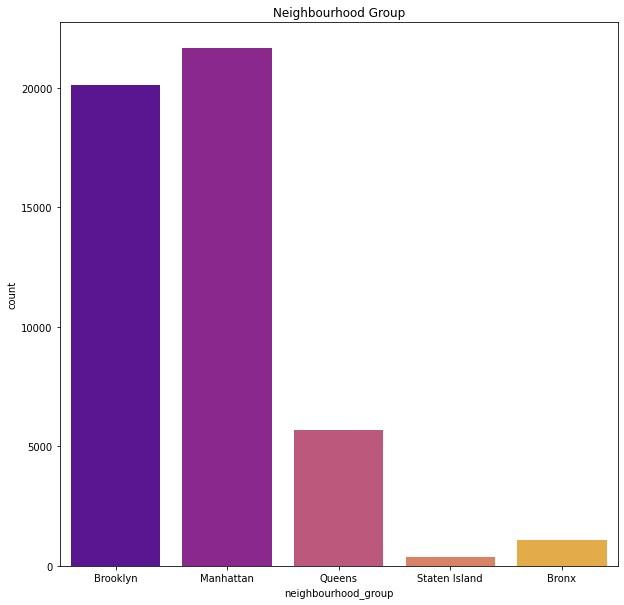
***dtype: int64***

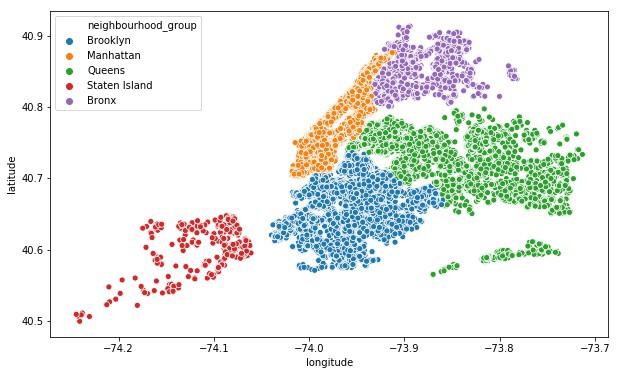
اطلاعات از دست رفته ای که مشاهده می شوند به راهکار خاص زیادی احتیاج ندارند. با نگاهی به ماهیت مجموعه داده های خود می توانیم موارد دیگری را بیان کنیم: ستون های "نام" و "نام میزبان" برای تجزیه و تحلیل داده های ما بی ربط و ناچیز هستند ، ستون های "last\_review" و "review\_per\_month" به مدیریت بسیار ساده نیاز دارند. برای توضیح بیشتر ، "last\_review" تاریخ است. برای ستون "review\_per\_month" می توانیم آن را به سادگی با 0.0 برای مقادیر از دست رفته ضمیمه کنیم. لطفا توجه داشته باشید که ما 'host\_name' را به دلیل ناچیز بودن و دلایل اخلاقی نیز رها می کنیم. برای ادامه اکتشاف داده ها و آموزش مدل (که بعداً انجام خواهیم داد) نسبت به افراد خاص بر اساس نام آنها هیچ استدلالی وجود ندارد. این اسامی به انسان واقعی اختصاص داده می شوند.

کاوش داده ها با تجزیه و تحلیل آمار و تجسم مقادیر ویژگی ها و همبستگی بین ویژگی های مختلف.

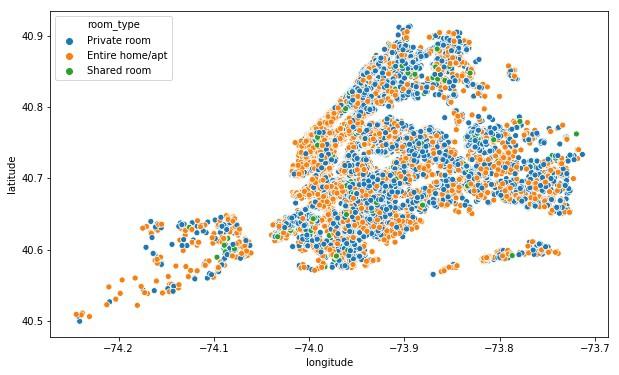


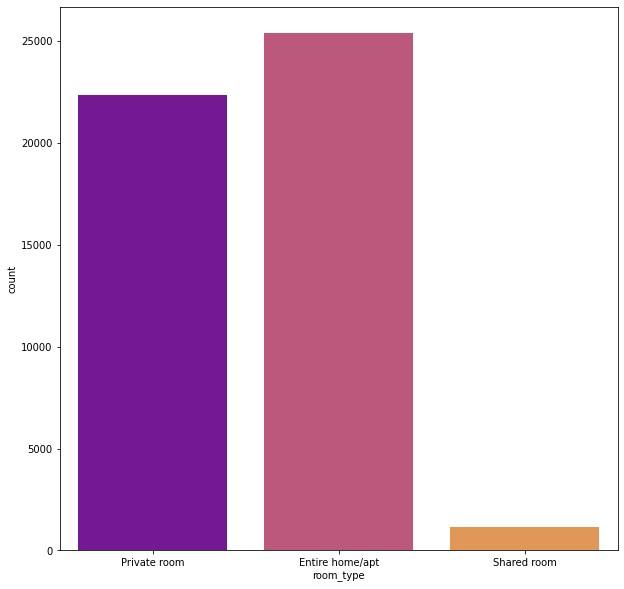
با توجه به نمودار ستونی می بینیم توزیع خوبی بین 10 میزبان برتر با بیشترین لیست وجود دارد. میزبان اول بیش از 300 لیست دارد.

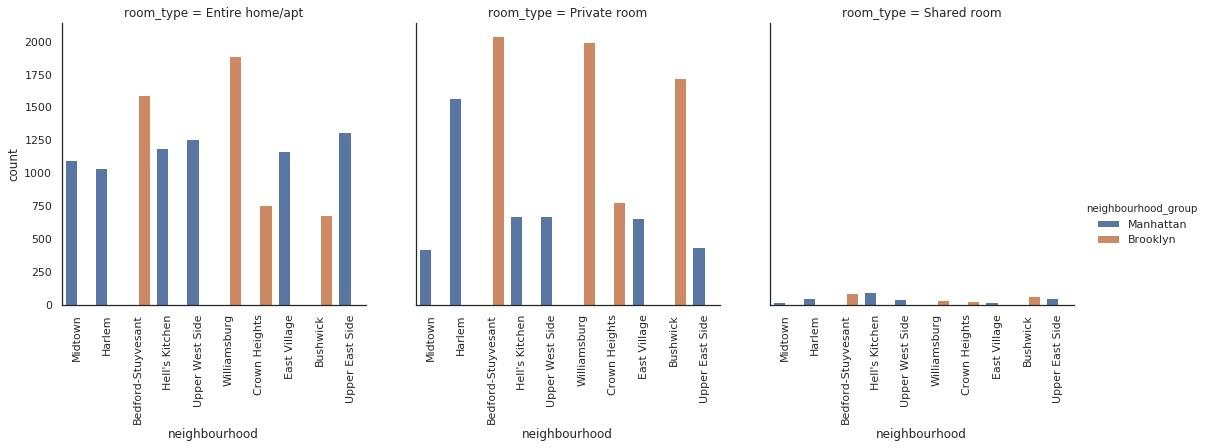




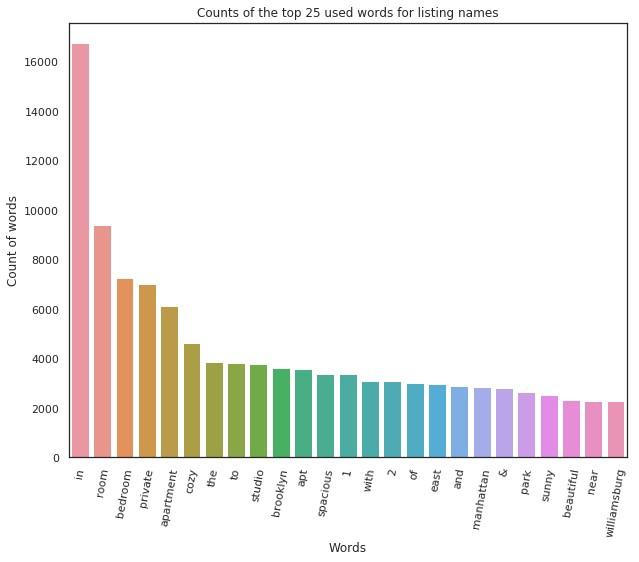
با داشتن یک جدول آماری و یک طرح ویولن ، ما قطعاً می توانیم چند مورد را در مورد توزیع قیمت Airbnb در بخش های NYC مشاهده کنیم. اول ، می توانیم بگوییم که منهتن بالاترین طیف قیمت را برای لیست ها با قیمت 150 دلار به عنوان مشاهده متوسط ​​دارد و پس از آن بروکلین با 90 دلار برای هر شب. کوینز و استاتن آیلند توزیع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



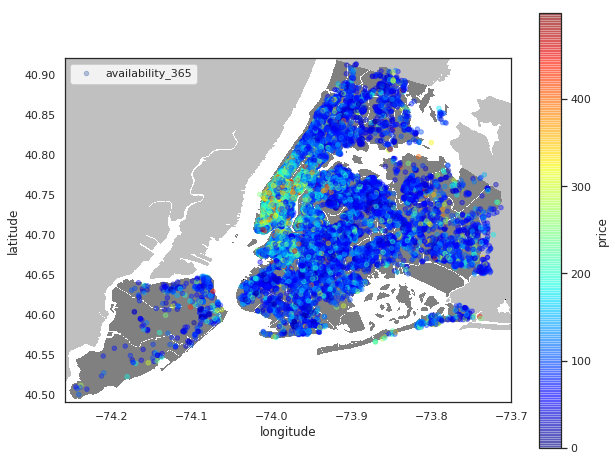


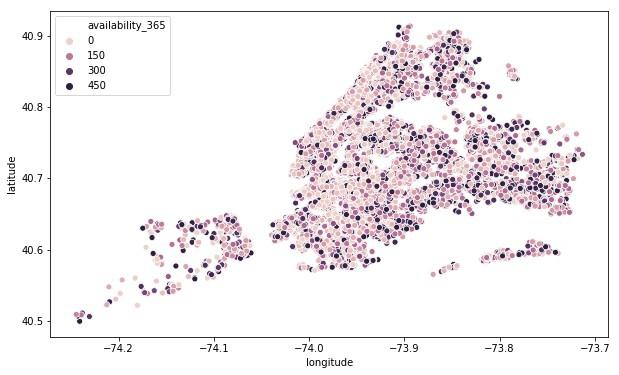


با توجه به نمودار های بالا اول ، می توانیم ببینیم که طرح ما از 3 زیرمجموعه تشکیل شده است - با چنین خروجی ، ما به راحتی می توانیم مقایسه توزیع بین ویژگی ها را ادامه دهیم. محورهای Y و X برای هر زیر قطعه کاملاً یکسان باقی می مانند ، محور Y نمایانگر شماری از مشاهدات محور X است که می خواهیم آنها را بشماریم. با این حال ، 2 عنصر مهم دیگر وجود دارد: ستون و رنگ. پس از تعیین ستون و تعیین رنگ ، قادر به مشاهده و مقایسه محورهای Y و X خود در میان ستون مشخص شده و همچنین کدگذاری رنگی هستیم. مشاهداتی که قطعاً بیشترین تضاد را دارند این است که لیست Airbnb از نوع "اتاق مشترک" 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.

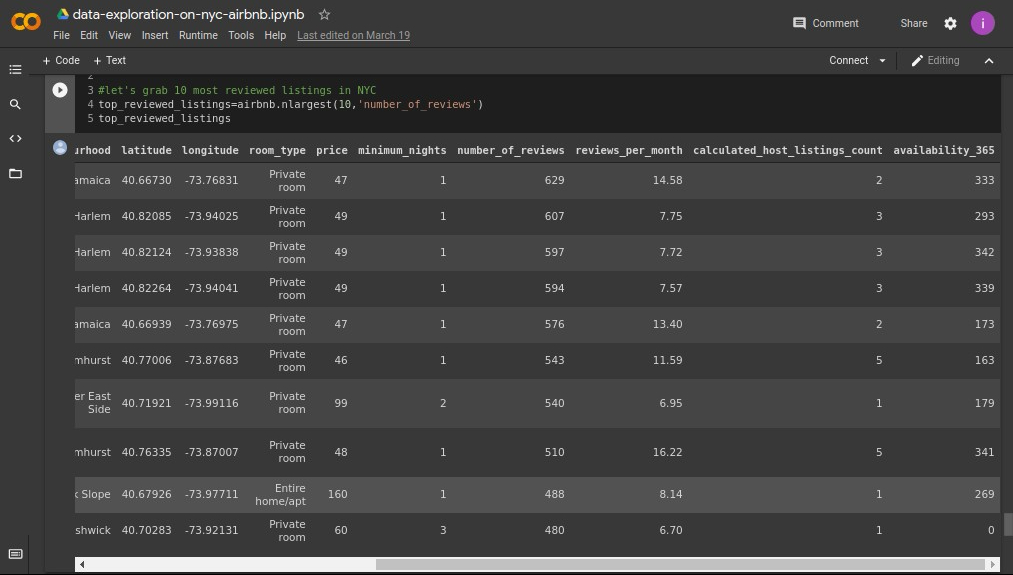


ما می توانیم با این نمودار ستونی 25 کلمه نامی لیست های پرکاربرد را کشف و مرور می کنیم . این نشان می دهد که میزبان ها به سادگی لیست خود را به صورت کوتاه با اصطلاحات بسیار خاص برای جستجوی راحت تر توسط یک مسافر بالقوه توصیف می کنند. چنین شرایطی عبارتند از "اتاق" ، "اتاق خواب" ، "خصوصی" ، "آپارتمان" . این نشان می دهد که هیچ عبارتی یا اصطلاحی "محبوب / پرطرفدار" وجود ندارد که برای نام استفاده شود. میزبان از اصطلاحات بسیار ساده ای برای توصیف فضا و منطقه ای که لیست در آن قرار دارد استفاده می کنند. این روش تا حدودی مورد انتظار بود زیرا برخورد با مشتریان چند زبانه می تواند مشکل باشد و شما قطعاً می خواهید فضای خود را تا حد ممکن به صورت مختصر و قابل فهم توصیف کنید.





در این عکس ها با استفاده از نقاط عرض و طول تمام لیست های NYC را تجسم کردیم. همچنین ، بر اساس قیمت لیست ، برای هر نقطه از نقشه محدوده کدگذاری رنگی اضافه کرده ایم.

****

از این جدول می توان مشاهده کرد که 10 لیست مورد بررسی بیشتر در Airbnb برای NYC دارای قیمت متوسط ​​ 65 $ با بیشتر لیست ها زیر $ 50 و 10/9 آنها از نوع "اتاق خصوصی" هستند. لیست برتر دارای 629 بررسی (review)است.

نتیجه گیری

به نظر می رسد این مجموعه داده برای سال 2019 یک مجموعه داده بسیار غنی با ستون های متنوع است که به ما امکان می دهد کاوش عمیق داده ها را در هر ستون قابل توجه ارائه دهیم. اول ، ما میزبانانی پیدا کرده ایم که از پلت فرم Airbnb به خوبی بهره می برند و بیشترین لیست را ارائه می دهند. ما دریافتیم که میزبان برتر ما 327 لیست دارد. پس از آن ، ما به تحلیل محله ها و تراکم لیست محله و مناطق محبوب تر از مناطق دیگر پرداختیم. در مرحله بعد ، ما از ستون های عرض و طول جغرافیایی خود برای ایجاد یک نقشه حرارتی جغرافیایی با کد رنگی با قیمت لیست ها استفاده می کنیم. علاوه بر این ، ما به ستون اول با رشته های نام بازگشتیم و مجبور شدیم برای تجزیه هر عنوان و تجزیه و تحلیل روند موجود در مورد نحوه نامگذاری لیست ها و همچنین تعداد کلمات مورد استفاده توسط میزبان ، کمی بیشتر کدگذاری کنیم. سرانجام ، بیشترین میزبان بازبینی شده را پیدا کردیم و برخی از ویژگی های اضافی را تجزیه و تحلیل کردیم. این تجزیه و تحلیل داده ها برای تصمیم گیری بهتر در زمینه کسب و کار ، کنترل بر روی سیستم ، ابتکارات بازاریابی ، اجرای ویژگی های جدید و موارد دیگر استفاده می شود.