داده کاوی

پروژه اول

دیتاست اول

دکتر فراهانی – دکتر شریفی

ابراهیم شکیباپور

۹۹۴۲۲۱۱۷

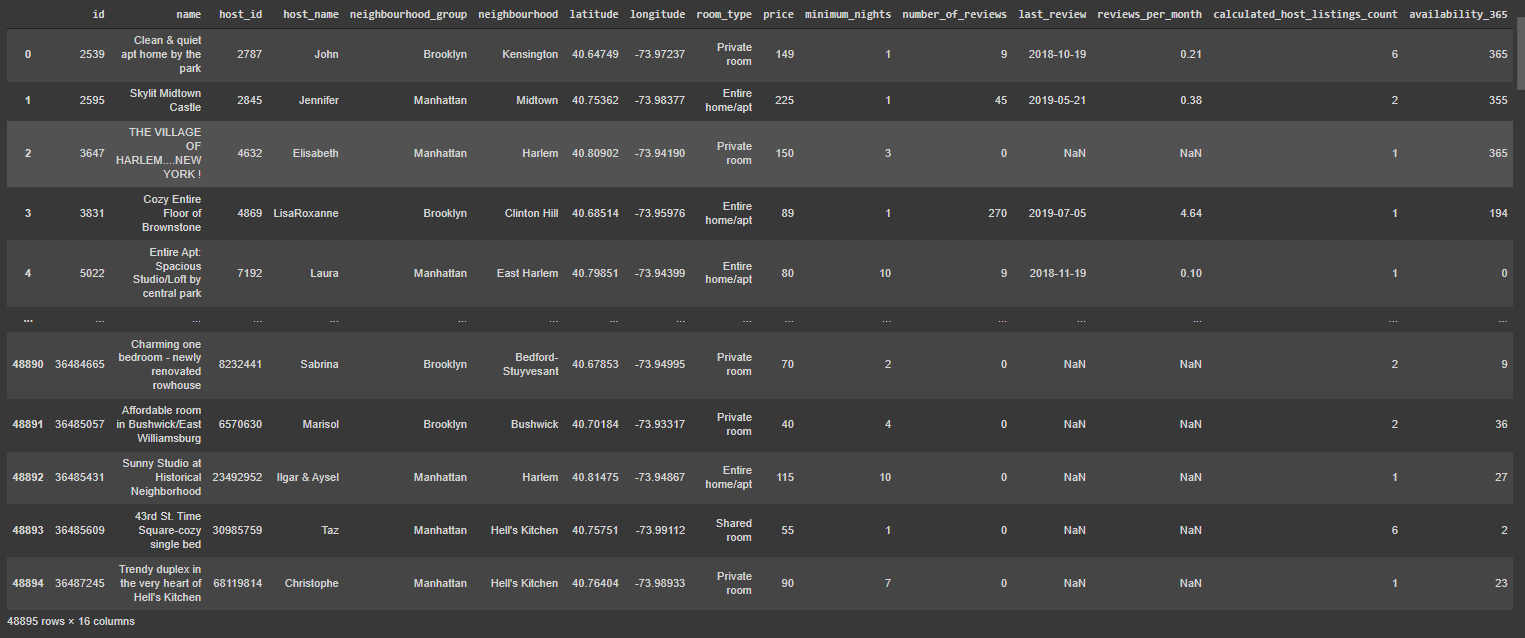
درباره ی دیتاست

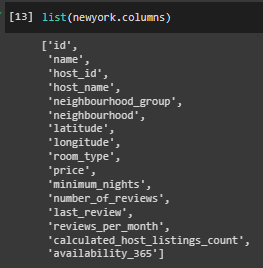
از سال 2008، مهمانان و میزبانان از Airbnb برای گسترش امکانات سفر و ارائه روش منحصربفردتر و شخصی‌شده‌تر برای تجربه جهان استفاده کرده‌اند.

این مجموعه داده فعالیت و معیارهای فهرست را در نیویورک، نیویورک برای سال 2019 توصیف می کند. این فایل داده شامل تمام اطلاعات مورد نیاز برای کسب اطلاعات بیشتر در مورد میزبان ها، در دسترس بودن جغرافیایی، معیارهای لازم برای پیش بینی و نتیجه گیری است.

بررسی و تحلیل دیتاست

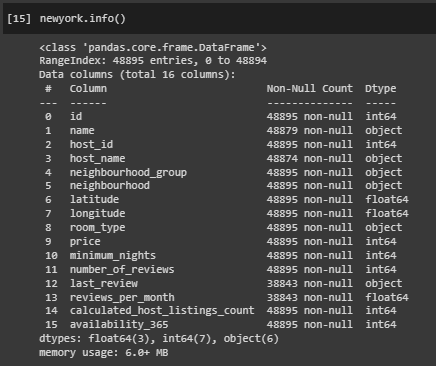
در گام نخست می توانیم ببینیم که دیتاست ما شامل ۴۸۸۹۵ سطر و ۱۶ ستون می باشد.



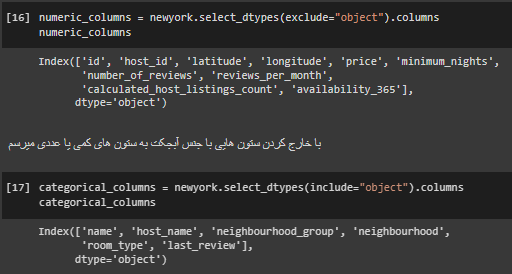


عناوین ستون ها را می توانیم در شکل بالا مشاهده کنیم

که بر اساس ماهیت دیتاست می تواند اطلاعات مفیدی به ما بدهد.

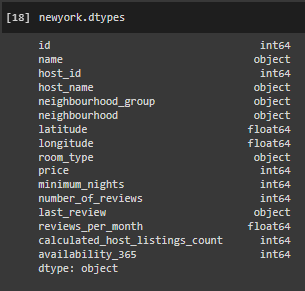


نوع داده مربوط به هر ستون را نیز مشاهده می کنیم

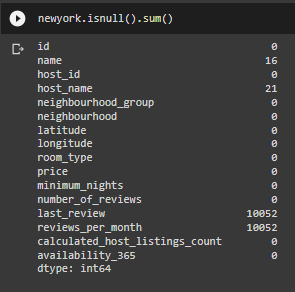


با خارج کردن ستون های با جنس آبجکت می توانیم ستون های کمی یا عددی را ببینیم

و همینطور اگر ستون هایی که آبجکت هستند را بررسی کنیم اطلاعات فوق را خواهیم داشت که بیانگر ستون های کیفی است

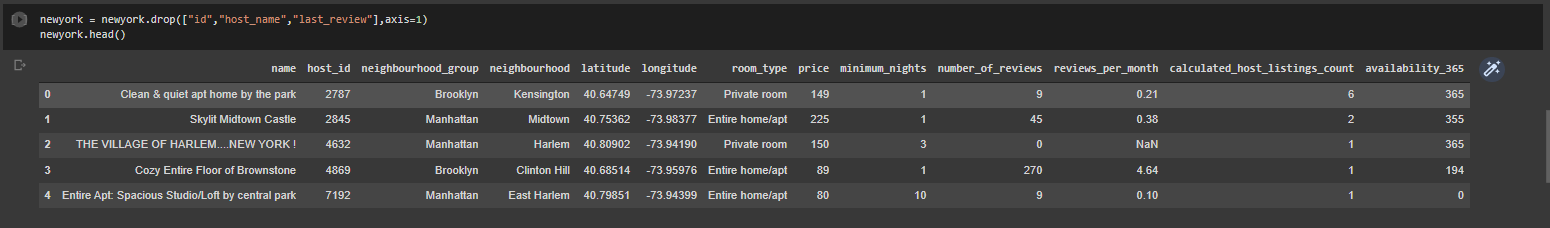


در نتیجه ده ستون دیتاست ماهیت کمی یا عددی و تعداد شش ستون نیز مقادیر کیفی دارند

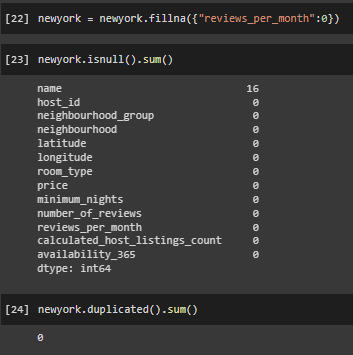


مشاهده می کنیم که دیتاست شامل داده ها و سلول های فاقد مقدار هست که این مقادیر می توانند حذف شوند و یا جایگزین مقادیر دیگر شوند

برخی از ستون ها با توجه به هدفی که داریم و همچینین ماهیت دیتاست اهمیت چندانی ندارند

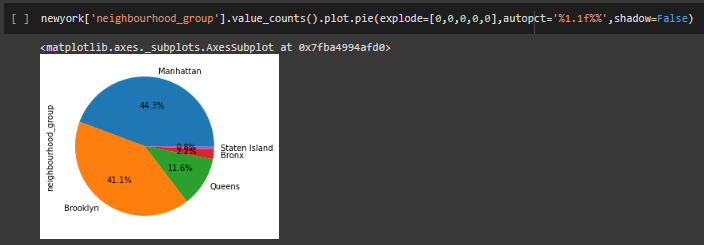


مقادیر آیدی, نام میزبان و آخرین بازدید را حذف کردیم

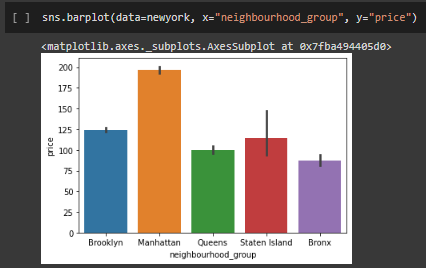


و داده های فاقد مقدار ستون تعداد بازدید هر ماه را برابر صفر قرار دادیم

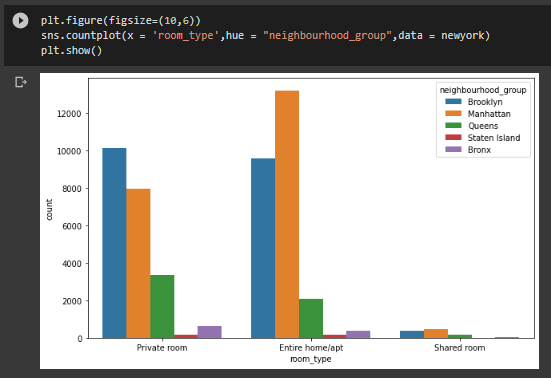
و همچنان مشاهده می کنیم مقادیر تکثیر شده و یا تکراری هم نداریم.



تعداد آگهی های مربوط به هر ناحیه از نیویورک را نیز در پای چارت فوق می بینیم که به ترتیب منهتن با ۴۴/۳ درصد, بروکلین با ۴۱/۱ درصد, کوئینز با ۱۱/۶ درصد و ...دارای بیشترین آگهی ها می باشند

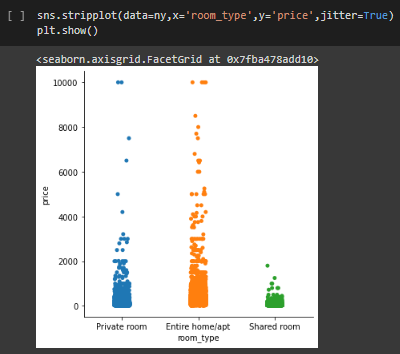


و همینطور در مورد قیمت هم منهتن و بروکلین دارای بیشترین قیمت ها هستند



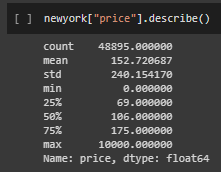
در بررسی نوع اتاق هم می توانیم ببینیم هر منطقه بیشتر دارای چه نوع اتاقی است

که می تواند ارتباط مستقیمی با قیمت خانه در هر منطقه داشته باشد

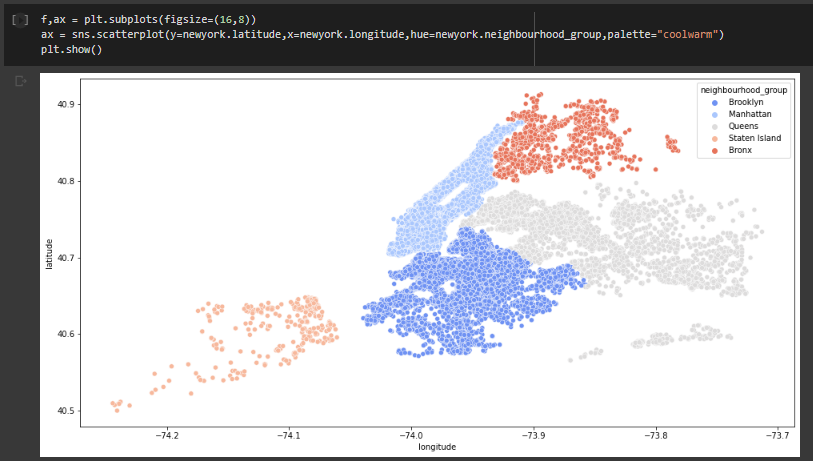


ازه های قیمتی مربوط به نوع اتاق نیز دارای گستره ی بالایی است

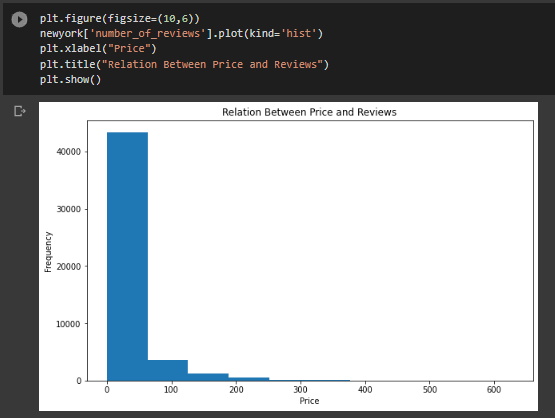
ولی تراکم قیمتی را در پایین نمودار می بینیم



شمای کلی از اطلاعات مربوط به قیمت ملک در نیویورک من جمله تعداد, میانگین قیمتی, انحراف معیار و ...



تراکم و توزیع آگهی دهندگان و موقعیت ملک ها در طول و عرض جغرافیایی



رابطه بین قیمت و دیدگاه ها یا نظرات را نیز در نمودار بالا می بینیم

هر چه قیمت پایین تر بوده دیدگاه ها و نظرات نیز بیشتر بوده است



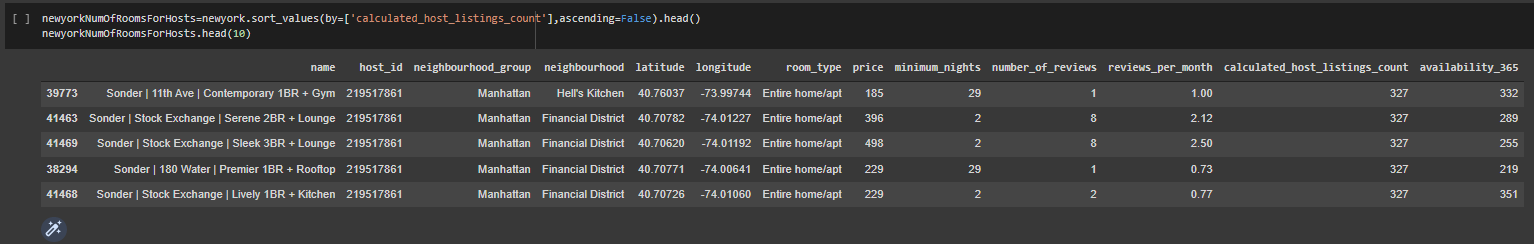
صاحبان آگهی که دارای بیشترین بازدید می باشند

و تعداد بازید زیاد همانطور که دیدم با قیمت کم رابطه مستقیم دارد.

یکی از دلایل مرتبط با مشتری زیاد می تواند ناحیه که خانه ها در آن قرار دارند باشد همینطور که می بینیم همه ی این آگهی ها در یک طول و عرض جغرافیایی خاصی قرار دارند.

نوع خانه شخصی می باشد.

دارای کمترین قیمت می باشند.



گزارش مربوط به صاحبان آگهی که دارای بیشترین تعداد آگهی می باشند