به نام خدا

تحلیل داده های وضعیت مسکن در شهر پکن چین با تمرکز بر تاثیر فاصله جغرافیایی هر خانه روی قیمت آن

اساتید:

مریم بابایی

متین جابری

سپهر رضایی

پژوهشگران:

طاها رحمانی[1]

حنانه پنجه‌خامنه[2]

[1] دانشجوی رشته آمار مقطع کارشناسی دانشگاه شهید بهشتی

[2] دانشجوی رشته آمار مقطع کارشناسی دانشگاه شهید بهشتی

مقدمه

دیتاهای در دست مربوط به خانه هایی در شهر پکن است که در رابطه با اطلاعات نمونه مورد بررسی، یک دیتاست دارای ۳۱۸۸۵۱ دیتا است که البته ۳۱۸۸۱۹ تعداد از دیتاها در اصل مورد بررسی قرار داده شده است. این دیتاها در ۱۸ ستون هستند که در هر ستون یک ویژگی از این دیتاها نظیر قیمت خانه، تعداد طبقات، سال ساخت خانه و ... نشان داده است. دیتاهای مورد بررسی ابتدا به فرمتی که کاراکترهای آن قابل خواندن باشند تبدیل شده اندو سپس اطلاعات هر ردیف بررسی شده اند و بعد از مواجه شدن با دیتاهای گم‌شده، با مد جایگزین شد و بعد هم داده های پرت از میان دیگر داده ها حذف شد. با توجه به اطلاعات موجود دو مشخصه قیمت متراژ خانه‌ها و فاصله خانه تا مرکز شهر پکن حاصل شدند. همبستگی و ارتباط بین آنها بررسی شد و درنهایت انتظار می‌رود که خانه‌هایی که مرک شهر نزدیک‌تر هستند قیمت بیشتری دارند.

روش تحقیق

در اولین گام، دیتاست بطور کلی بررسی شد و پس از شناسنایی مشخصه‌های غیرضروری فیلتر شدند. باتوجه به این که پروژه حاضر یک دیتاست موثق از خانه های شهر پکن است بعضی از داده ها به زبان چینی نوشته شده‌اند؛ به همین دلیل فایل csv با انکودر gbk خوانده شد و در housing ذخیره شد.

ویژگی های ستون های دیتاها

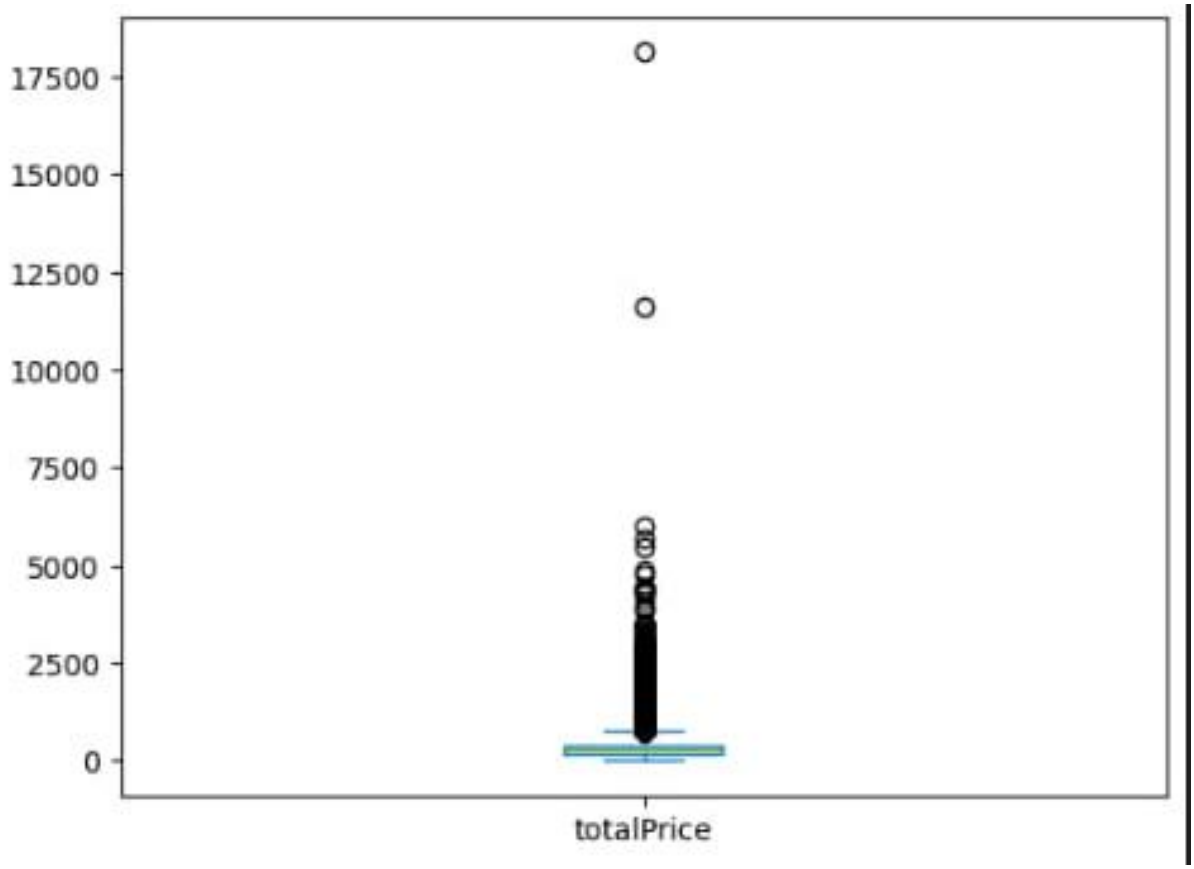
Unnamed: 0	نمایه ی داده هاست که نیازی به آن نداریم و باید حذف شود.
url	لینک معامله
id	شناسه معامله
Lng	طول جغرافیایی خانه معامله شده
Lat	عرض جغرافیایی خانه معامله شده
Cid	شناسه خریدار
tradeTime	زمان انجام معامله
DOM	اعداد روزی که از زمان گذاشتن آگهی خانه میگذرد.
totalprice	قیمتی که خانه به فروش رفته است.
Square	متراژ خانه
LivingRoom	تعداد اتاق نشیمن
drawingRoom	تعداد اتاق پذیرایی
Kitchen	تعداد آشپزخانه
bathroom	تعداد حمام
floor	طبقه و ارتفاع خانه
constructionTime	سال ساخت خانه
renovationCondition	وضعیت نوسازی خانه
buildingStructure	ساختار خانه
ladderRatio	طبقه و ارتفاع خانه
elevator	خانه آسانسور داردد یا خیر
subway	خانه به مترو دسترسی دارد یا خیر
district	منطقه ای که خانه در آن قرار دارد

سه ستون id ,url ,Cid و ستون اول که نمایه داده‌ها بود به دلیل ناکارآمدی در این پژوهش حذف شدند و دیتافرم جدید بدون ستون‌های اضافه در دیتافریم housing_dropped ذخیره شد. برای بررسی بیشتر دیتاست و تصمیم‌گیری برای داده‌های گم‌شده، مشخصه‌های مختلف و تعداد داده‌های گمشده در آن مشخصه را در یک دیتافریم ذخیره کردیم.

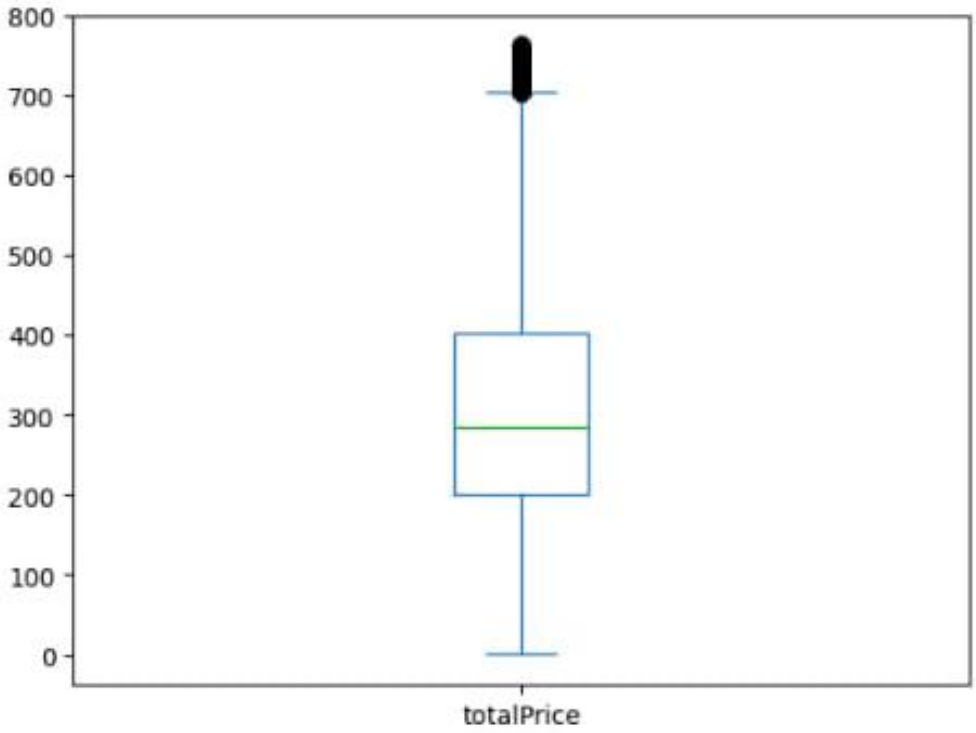
جدول ویژگی‌های دیتاست و تعداد داده‌های گمشده

Lng	۰
Lat	۰
tradeTime	۰
DOM	۱۵۷۹۷۷
totalprice	۰
Square	۰
LivingRoom	۰
drawingRoom	۰
Kitchen	۰
bathroom	۰
floor	۰
constructionTime	۰
renovationCondition	۰
buildingStructure	۰
ladderRatio	۰
elevator	۳۲
subway	۳۲
district	۰

ستون DOM که نشان‌دهنده تعداد روزهای قرار گرفتن آگهی بر روی سایت تا روز فروش است، مقادیر گم شده زیادی دارد و اگر ردیف هایی که مقدار DOM آنها موجود نیست حذف شوند، تقریبا نیمی از داده ها حذف می‌شوند که در تحلیل داده‌ها تاثیرگذار است. نمودار جعبه‌ای این ستون به شکل زیر است:



واضح است که نمودار جعبه‌ای آن بسیار نامتقارن است و داده‌های پرت بسیاری در آن وجود دارد که نشان‌دهنده وجود خانه‌های بسیار گران یا بسیار بزرگ است. بنابراین استفاده از میانگین برای پر کردن داده‌های گمشده روش مناسبی نیست و بهتر است از مد استفاده شود زیرا تعداد زیادی از داده‌ها در یک بازه مشخصی(۰ تا ۷۵۰۰) قرار دارند. بنابراین داده‌های گمشده ستون DOM را با استفاده از مد همین ستون جایگذاری کردیم. پس از این فرایند نمودار جعبه‌ای این ستون به شکل زیر حاصل شد:



بعد از پاکسازی ستون DOM ستون‌های elevator و subway را بررسی کردیم که هرکدام ۳۲ داده گمشده دارند. تعداد این داده‌ها در بین تقریبا ۳۰۰هزار داده ناچیز است و پاک کردن این ردیف‌ها تاثیر زیادی روی نتایج کلی ندارند. بنابراین این داده‌ها را بطوری کلی حذف می‌کنیم و درنهایت یک دیتاست پاکسازی‌شده بدون دیتاهای گمشده و ناکارامد داریم. دیتافریم پاکسازی‌شده را در یک دیتافریم جدید به نام housing_no_missing ذخیره می‌کنیم.

با بررسی بیشتر دیتاست روشن است تعدادی از ستون ها دارای مقادیر categorical هستند مانند constructionTime و floor که برای نمایش بهتر داده‌ها این اعداد به معادل رشته‌ای آنها تبدیل شد.

مقادیر ستون‌های renovationCondition, elevator, subway مطابق مقادیر زیر تغییر داده شدند و در یک دیتافریم جدید به نام housing_categorical ذخیره شد.

subway

Has subway	1
no subway	0

elevator

has elevator	1
no elevator	0

علاوه‌بر موارد فوق اطلاعاتی که در ستون‌های سال ساخت خانه و ارتفاع خانه هستند برخلاف ماهیت عددی مقداری رشته‌ای دارند که باید به ماهیت عددی تبدیل شوند. بااستفاده از متد unique() مقادیر یکتای هرکدام از ستون‌ها را مشخص کردیم. مشاهده می‌شود که در ستون constructionTime بعضی از خانه‌ها با کلمات چینی به معنای نامشخص "unknown" پر شده که چون تعدادشان کم است حذف شدند و پس از آن نوع ستون constructionTime به int تغییر داده شد و در دیتافریم جدید به نام housing_construction ذخیره شد. ستون مورد بررسی بعدی ستون floor بوده است. به‌گونه‌ای که در خانه‌های این ستون ابتدا یک کلمه چینی و پس از آن یک عدد قرار گرفته است که عدد بعد از کاراکتر چینی ارتفاع خانه را نشان می‌دهد. پس از بررسی معنی کلمات روشن شد که کلمات به معنای بالا، پایین و متوسط بوده‌اند. ستون floor طوری تغییر داده شد که فقط اعداد آنها باقی ماند و جنس داده‌های این ستون با استفاده از توابع پانداز به int تبدیل شد و در دیتافریم جدید به نام housing_floor ذخیره شد.

در این پژوهش قرار است تاثیر فاصله جغرافیایی خانه‌ها تا مرکز شهر روی قیمت آن‌ها بررسی شود.

با استفاده از داده های موجود، داده های جدید به دست آورده شد مانند فاصله هر خانه تا مرکز پایتخت چین و قیمت معامله شده برای هر متر از خانه. داده های جدید از روی داده های موجود جهت تحلیل آماری بیشتر و بهتر مورد استفاده قرار گرفتند. فاصله هر خانه تا مرکز پایتخت چین محاسبه در ستونی به اسم

distanceToCapital ذخیره شد. طول و عرض جغرافیایی مرکز پایتخت چین به ترتیب ۱۱۶.۴۰۷۴ و ۳۹.۹۰۴۲ است که برای محاسبه فاصله هر خانه تا مرکز پایتخت چین از فرمولی که برای محاسبه فاصله دو نقطه جغرافیایی است استفاده شد که با استفاده از نامپای انجام شد.(R شعاع کره زمین است)

$$Distance = acos (siny1 * siny2 + cosy1 * cosy2 * cos (x2 – x1)) * R$$

$$y = lat$$

$$x = lng$$

$$R = 6371.0088$$

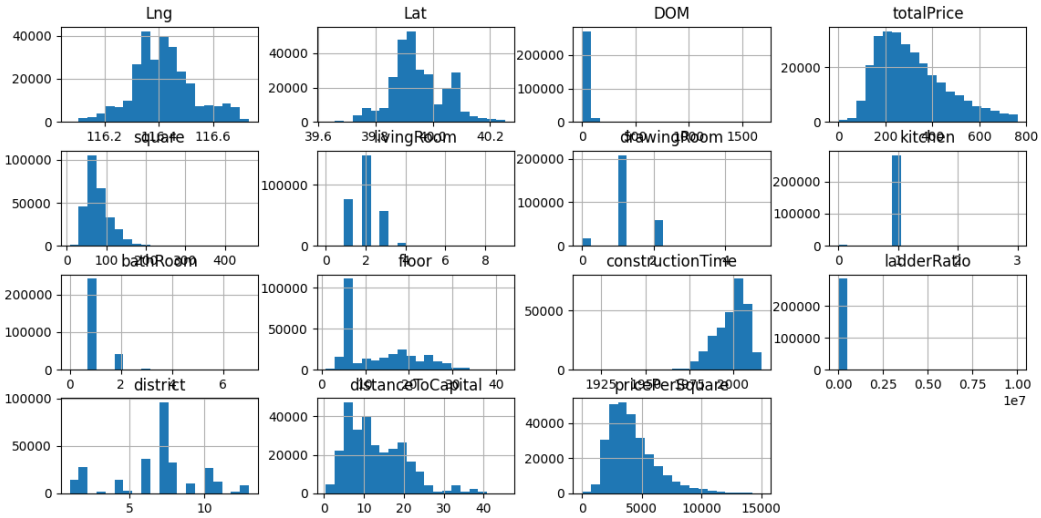
در این مجموعه داده، قیمت معامله شده و متراژ هر خانه مشخص است اما قیمت هر متر مربع خانه مشخص نیست. در ستون priceSquare قیمت هر متر مربع خانه محاسبه و ذخیره شد. قیمت کلی بر اساس میلیون یوان است اما قیمت هر متر مربع بر اساس یوان به دست آورده شد. دیتا فریم حاصل با نام housing_PPS ذخیره شد.

مجموعه این دو داده جدید نه داده پرت داشت و نه داده گمشده و نه با ستون هایی با فرمت نامناسب ذخیره شده بود. و اولین گام نگاهی به هیستوگرام ستون های عددی است. با توجه به نمودار، قیمت خانه به سمت قیمت ۲۰۰ میلیون یوان چوله شده است. اکثر خانه ها حوالی سال ۲۰۰۰ ساخته شده اند. بیشترین فاصله از مرکز پایتخت ٤٠ کیلومتر است. اکثر خانه‌ها دو اتاق نشیمن دارند و ... .

با استفاده از ستون جدید ساخته شده، میزان تغییر قیمت خانه با دور شدن از مرکز شهر بر روی یک نمودار نمایش داده شد.

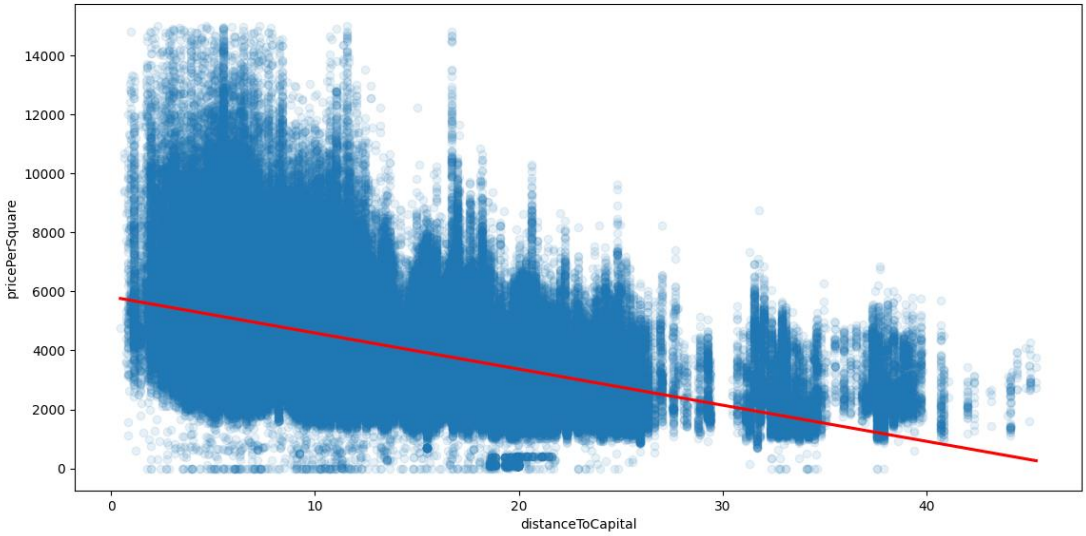
میزان اثر گذاری بودن یا نبودن آسانسور بر قیمت هر متر مربع خانه بررسی شد و برای این کار از نمودارهای مختلفی میشد استفاده کرد که نمودار به دست آمده در زیر نمایش داده شده است و دیتا فریم نهایی برای استفاده در مراحل بعدی housing_extended.csv ذخیره شد.

نمودار تمام ویژگی‌های خانه‌ها



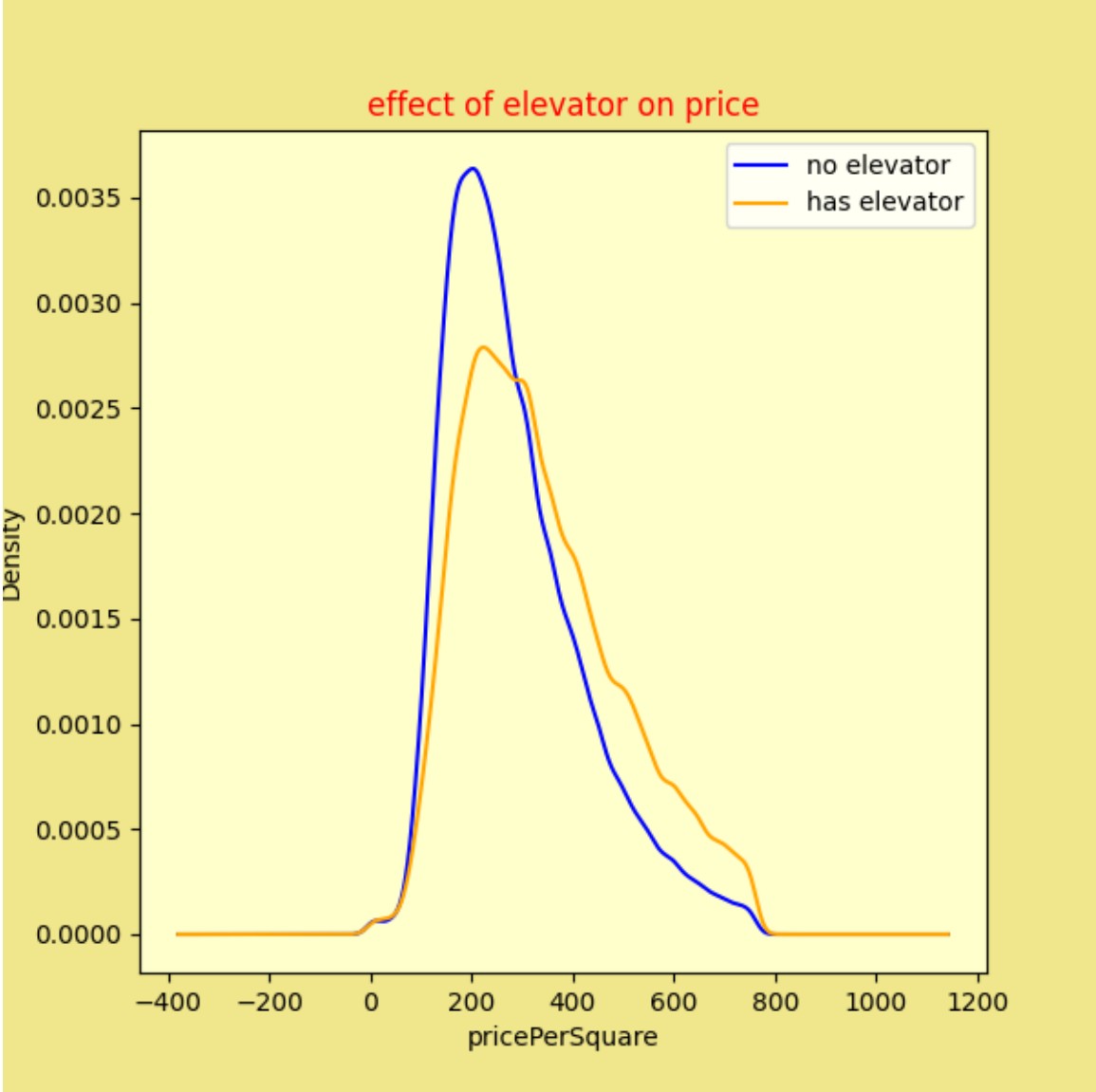
همچنین با توجه به نمودار فاصله تا مرکز شهر و قیمت خانه می‌توان نتیجه گرفت که هرچه موقعیت مکانی به مرکز شهر نزدیکتر باشد پس قیمت گرانتری هم دارد. نمودار زیر و خط رگرسیونی بیان‌گر این مفهوم است:

نمودار همسبتگی فاصله خانه از مرکز شهر و قیمت خانه‌ها



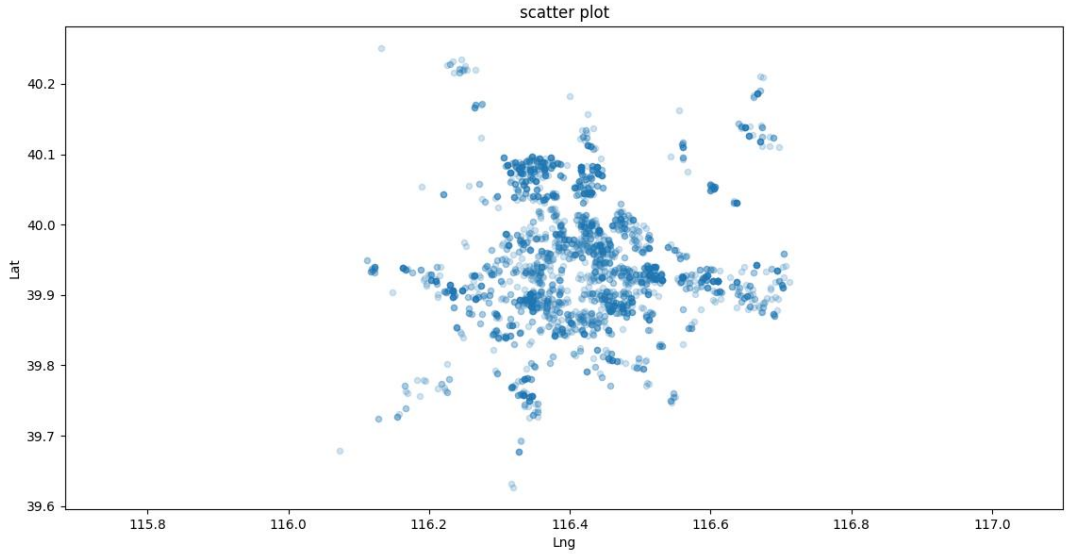
اکنون قصد داریم داده‌ها را به شکل دیگری تحلیل کنیم. میزان اثرگذاری داشتن یا نداشتن آسانسور روی قیمت خانه اهمیت دارد و می‌توان تصمیماتی بر مبنای آن گرفت. نمودار چگالی از مقایسه این دو ویژگی به این صورت می‌باشد:

تاثیر آسانسور روی قیمت خانه‌ها



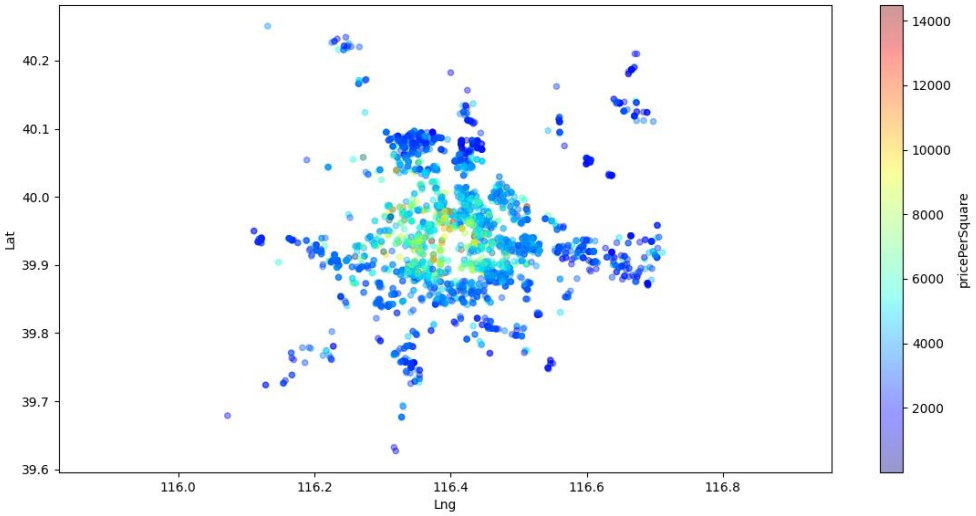
اکنون با تشکر به داشتن مختصات جغرافیایی هر خانه می‌توان خانه‌ها را بر اساس موقعیتشان روی نقشه رسم کرد. تراکم و پراکندگی و ویژگی‌های مختلف می‌توانند اطلاعات معناداری تولید کنند. نمودار scatter داده‌ها به شکل زیر است:

نمودار پراکندگی داده‌ها



اکنون با استفاده از تعیین رنگ هر نقطه قیمت خانه را مشخص می‌کنیم. به طوری که هرچی قیمت خانه بیشتر شود، رنگ آن نقطه به سمت رنگ‌های گرم‌تر متمایل می‌شود و هرچه قیمت خانه کمتر شود به سمت رنگ‌های سردتر متمایل می‌شود.

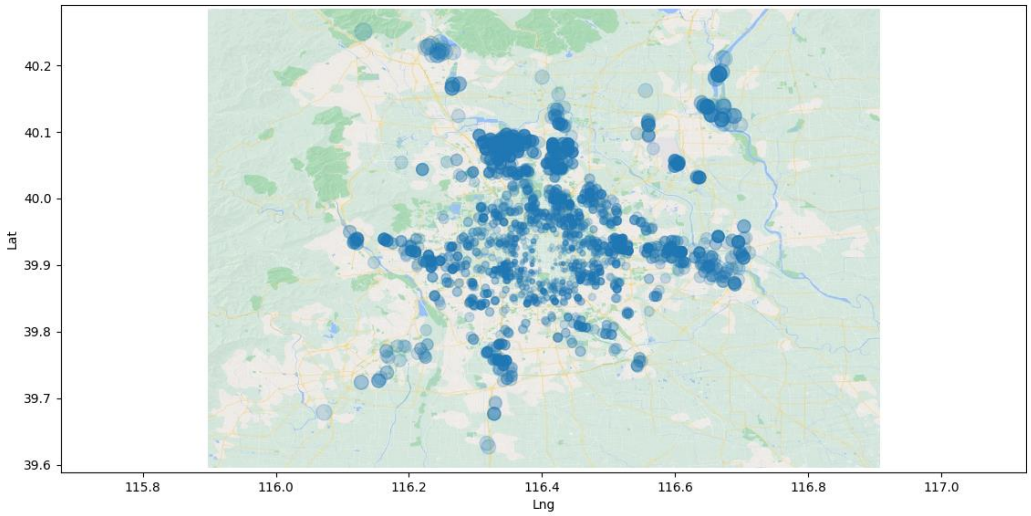
نمودار رنگی شده براساس تغییرات قیمت



واضح است که هرچه خانه‌ها به مرکز شهر نزدیکتر می‌شوند بر قیمت آن‌ها اضافه می‌شود و هرچه خانه‌ها به حاشیه شهر نزدیکتر می‌شوند از قیمت آن‌ها کاسته می‌شود. باتوجه به واقعی بودن موقعیت‌های مکانی می‌توانیم

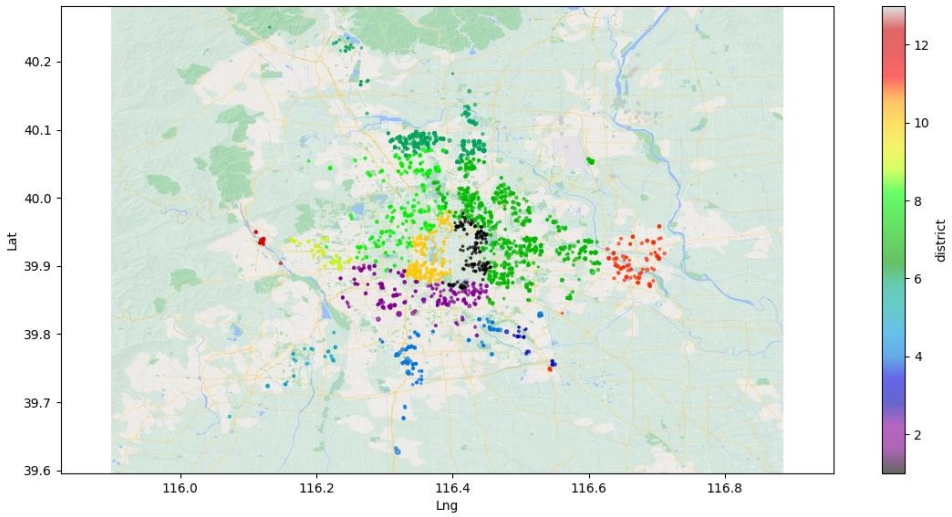
تصویر جغرافیایی شهر پکن را زیر نمودار تصویر کنیم. همچنین اندازه نقاط را معیاری برای ابعاد هر نقطه انتخاب کردیم که نمودار آن به شکل زیر است:

نمودار داده‌ها براساس فاصله از مرکز شهر و تصویر جغرافیایی



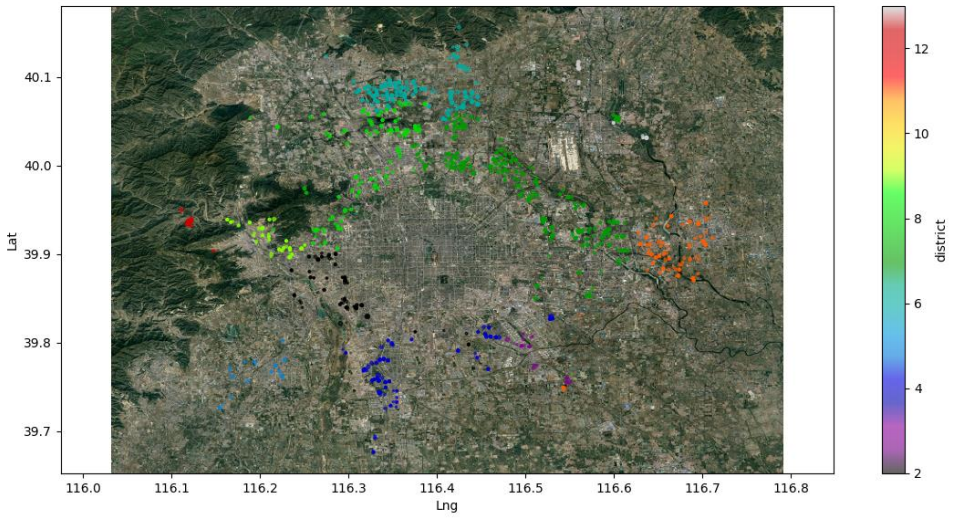
می‌توان نمودار دیگری براساس رنگ نقاط که نمایانگر قیمت خانه‌ها بود به شکل زیر ترسیم کرد:

نمودار داده‌ها بر اساس قیمت و موقعیت مکانی



نمودار فوق را می‌توان با تمرکز بر خانه‌هایی در فاصله ۱۰ تا ۳۰ کیلومتری از مرکز شهر هستند و استفاده از عکس واقعی‌تری از شهر پکن اصلاح کرد:

نمودار داده‌ها روی نقشه جغرافیایی و موقعیت مکانی خانه‌ها در فاصله ۱۰ تا ۳۰ کیلومتری



نتیجه‌گیری

باتوجه به نمودارهایی که رسم شد و تحلیل‌هایی که موقعیت مکانی خانه‌ها و وجود آسانسور در آن‌ها انجام شد انتظار می‌رود که قیمت خانه‌ها در مرکز شهر بیشتر است؛ بنابراین سرمایه‌گذاری روی مسکن در حاشیه شهر که قیمت پایین‌تری دارند مناسب‌تر خواهد بود. همچنین نمودار چگالی وجود یا عدم آسانسور نتیجه می‌دهد که در حاشیه شهر خانه‌های بدون آسانسور قیمت بالاتری دارند و در مرکز شهر داشتن آسانسور قیمت خانه را افزایش می‌دهد.

محدودیت‌های دیتاست

اگر وضعیت جنس ساختمان‌ها به طور دقیق مشخص بود و دسته‌بندی شده بود با توجه به این که پارامتری تاثیرگذار روی قیمت خانه است، این امکان را می‌داد که تحلیل‌هایی نظیر تاثیر مصالح ساختمانی روی قیمت خانه، وجود آسانسور در خانه‌هایی با مصالح ساختمانی خاص، جنس مصالح ساختمانی خانه‌های نزدیکتر به مترو و ... را ارائه داد.