

به نام خدا

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

اساتید:

مریم بابایی

متین جابری

سپهر رضایی

پژوهشگران:

طاها رحمانی<sup>۱</sup>

حنانه پنجه خامنه<sup>۲</sup>

---

<sup>۱</sup> دانشجوی رشته آمار مقطع کارشناسی دانشگاه شهید بهشتی

<sup>۲</sup> 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 اند و سپس اطلاعات هر ردیف بررسی شده اند و بعد از مواجه شدن با دیتاهای گم شده، 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 خانه ها و فاصله خانه تا مرکز شهر پکن حاصل شدند. همبستگی و ارتباط بین آنها بررسی شد و در نهایت انتظار می رود که خانه هایی که مرکز شهر نزدیک تر هستند قیمت بیشتری دارند.

## روش تحقیق

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

## ویژگی های ستون های دیتاها

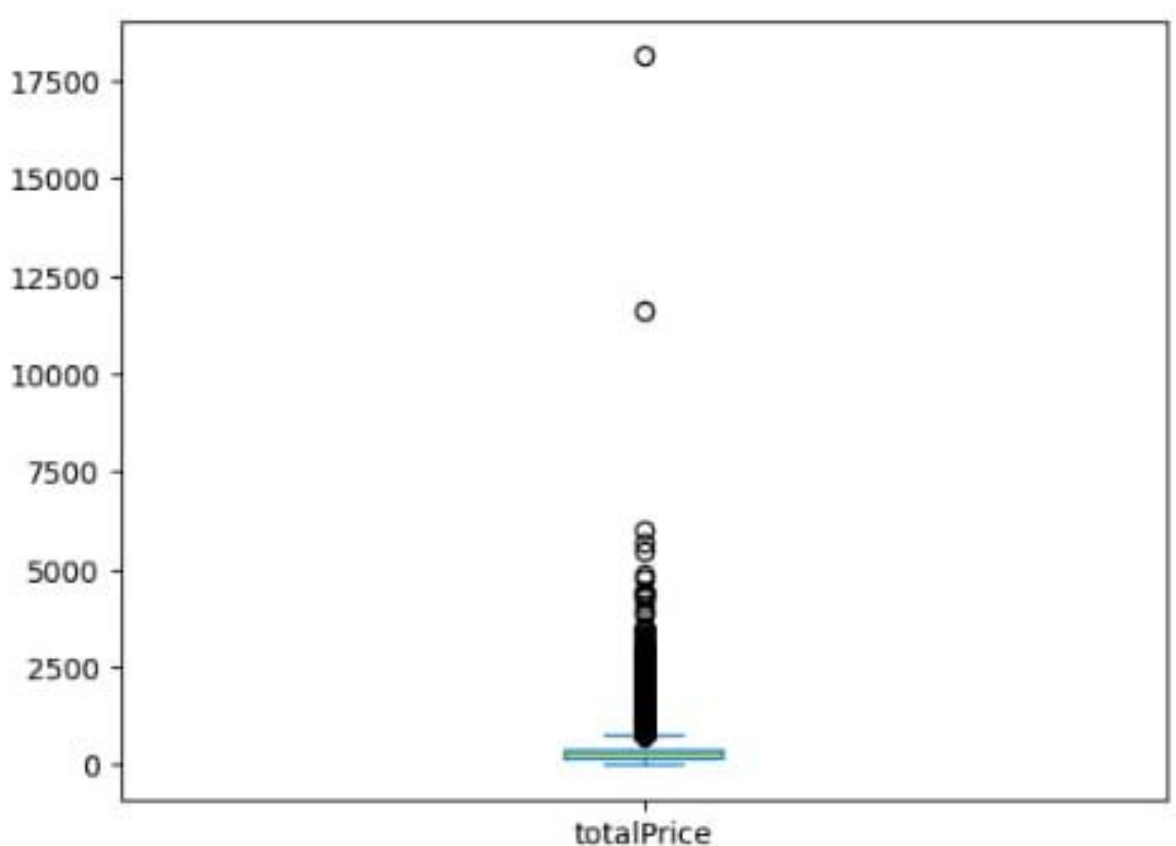
نمایه ی داده هاست که نیازی به آن نداریم و باید حذف شود.	Unnamed: 0
لینک معامله	url
شناسه معامله	id
طول جغرافیایی خانه معامله شده	Lng
عرض جغرافیایی خانه معامله شده	Lat
شناسه خریدار	Cid
زمان انجام معامله	tradeTime
اعداد روزی که از زمان گذاشتن آگهی خانه میگذرد.	DOM
قیمتی که خانه به فروش رفته است.	totalprice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

سه ستون Cid, url, id و ستون اول که نمایه داده‌ها بود به دلیل ناکارآمدی در این پژوهش حذف شدند و دیتافرم جدید بدون ستون‌های اضافه در دیتافرم 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 دیتافرم ذخیره کردیم.

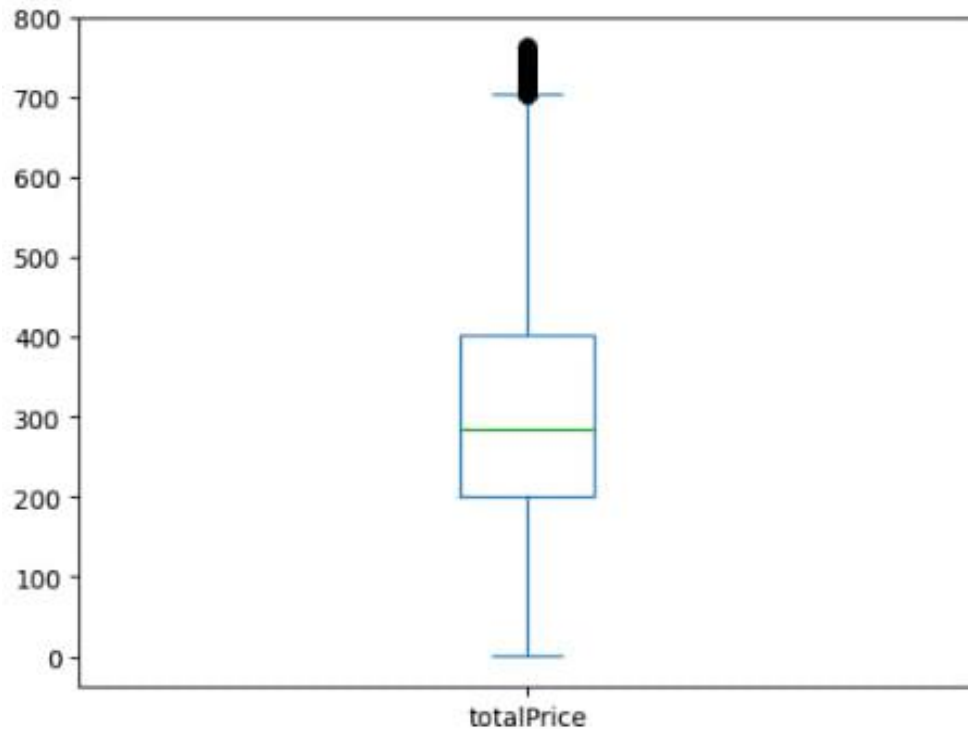
### جدول ویژگی‌های دیتاست و تعداد داده‌های گم‌شده

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 کم شده زیادی دارد و اگر ردیف‌هایی که مقدار DOM آنها موجود نیست حذف شوند، تقریباً نیمی از داده‌ها 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 تقریباً ۳۰۰ هزار داده ناچیز است و پاک کردن این ردیف‌ها تاثیر زیادی روی نتایج کلی ندارند. بنابراین این داده‌ها را بطوری کلی حذف می‌کنیم و در نهایت یک دیتاست پاکسازی شده بدون دیتاهای گمشده و ناکارآمد داریم. دیتافریم پاکسازی شده را در یک دیتافریم جدید به نام `housing_no_missing` ذخیره می‌کنیم.

با بررسی بیشتر دیتاست روشن است تعدادی از ستون‌ها 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

1	Has subway
0	no subway

### elevator

1	has elevator
0	no elevator

### renovationCondition

1	Other
2	Rough
3	Simplicity
4	hardcover

علاوه بر 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. با استفاده از متد `unique()` مقادیر یکتای هر کدام 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. به گونه‌ای 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 داده‌های موجود، داده‌های جدید به دست آورده شد مانند فاصله هر خانه تا مرکز پایتخت چین و قیمت معامله شده برای هر متر از خانه. داده‌های جدید از روی داده‌های 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)

$$\text{Distance} = \text{acos}(\sin y_1 * \sin y_2 + \cos y_1 * \cos y_2 * \cos(x_2 - x_1)) * R$$

$$y = \text{lat}$$

$$x = \text{lng}$$

$$R = 6371.0088$$

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.

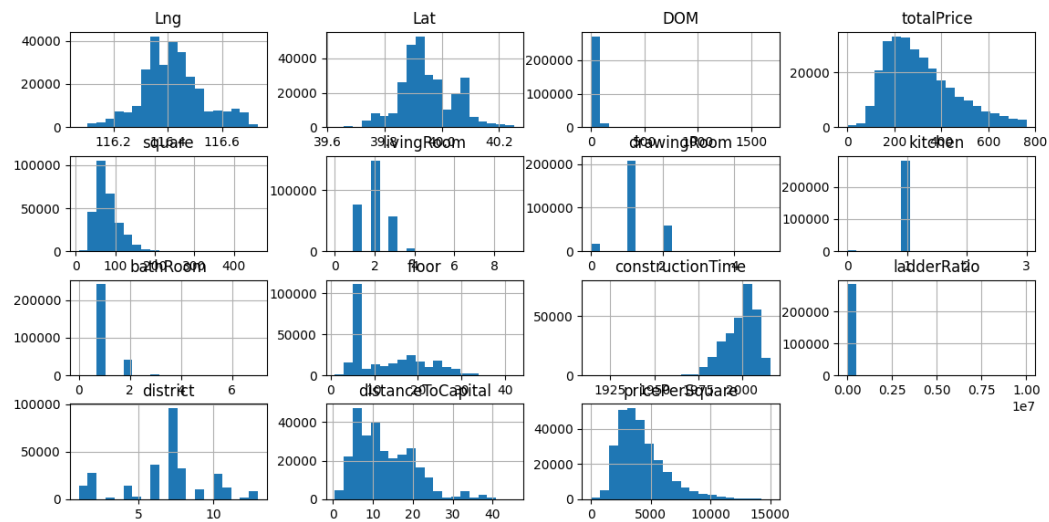
مجموعه این دو داده جدید نه داده پرت داشت و نه داده گم شده و نه با ستون‌هایی با فرمت نامناسب ذخیره شده بود. و اولین گام نگاهی به هیستوگرام ستون‌های عددی است. با توجه به نمودار، قیمت خانه به سمت قیمت ۲۰۰ میلیون یوان چوله شده است. اکثر خانه‌ها حوالی سال ۲۰۰۰ ساخته شده‌اند. بیشترین فاصله از مرکز پایتخت ۴۰ 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.



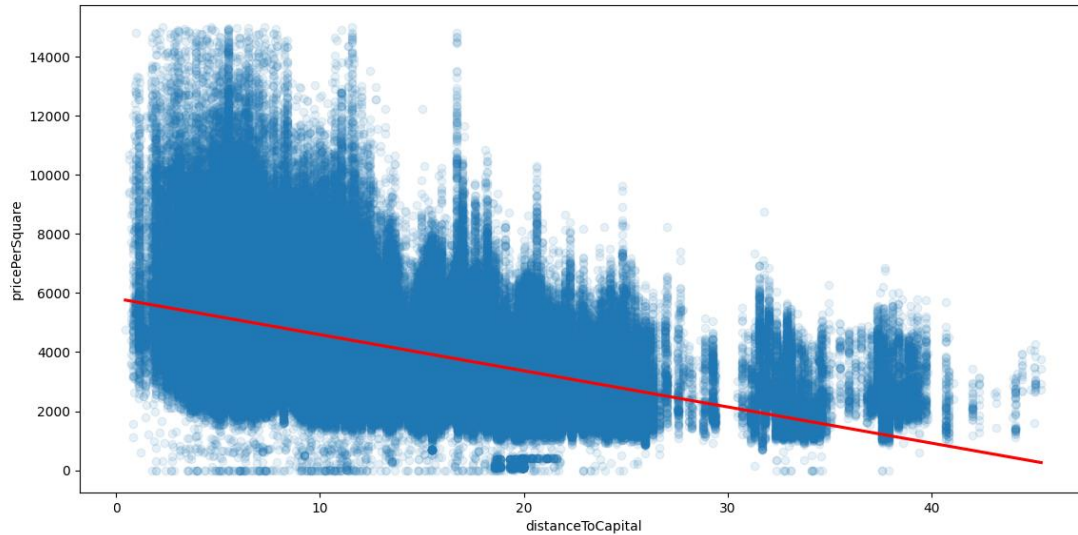
میزان اثر گذاری بودن یا نبودن آسانسور بر قیمت هر متر مربع خانه بررسی شد و برای این کار از نمودارهای مختلفی میشد استفاده کرد که نمودار به دست آمده در زیر نمایش داده شده است و دیتا فریم نهایی برای استفاده در مراحل بعدی housing\_extended.csv ذخیره شد.

## نمودار تمام ویژگی‌های خانه‌ها



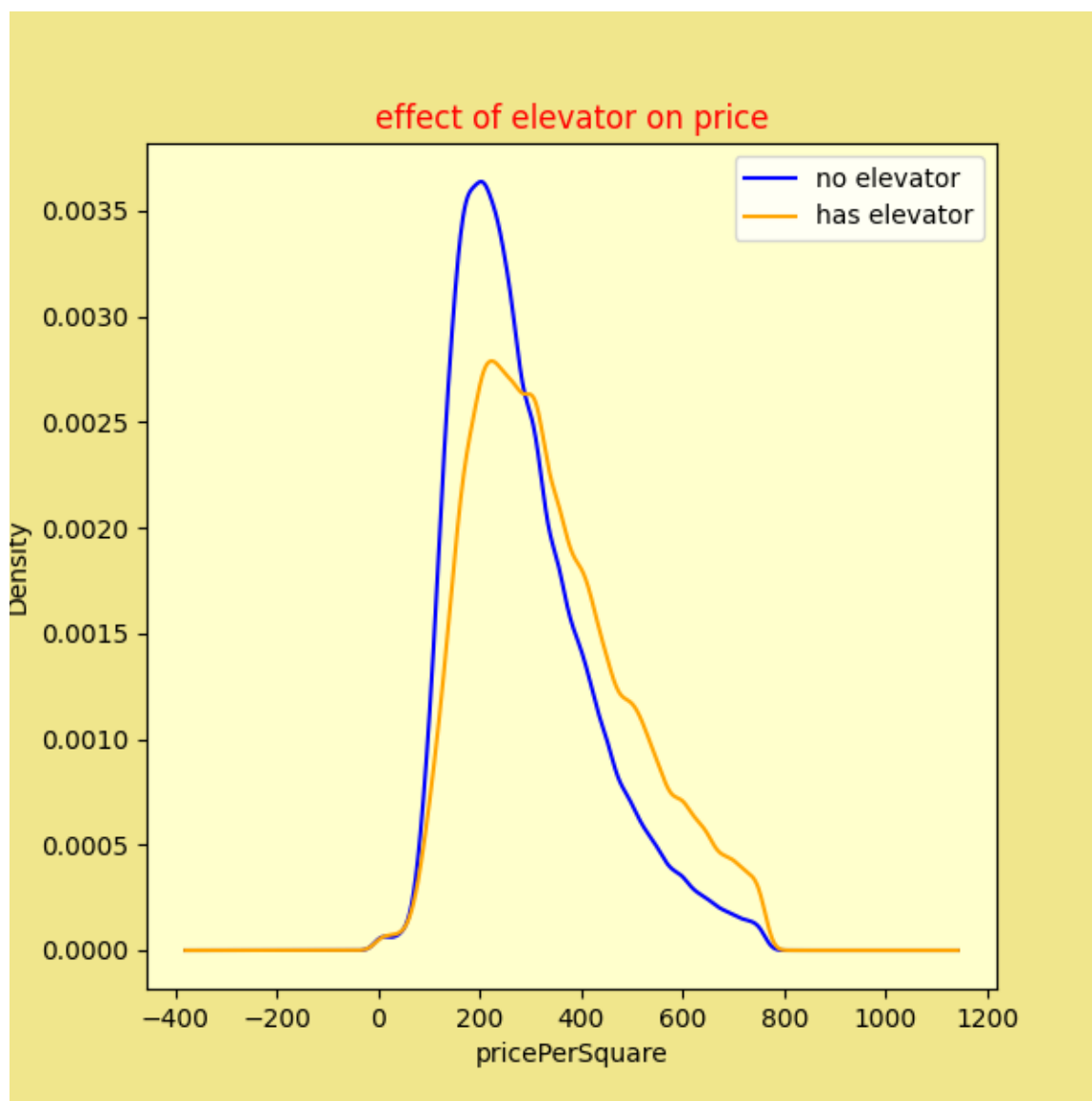
همچنین با توجه به نمودار فاصله تا مرکز شهر و قیمت خانه می‌توان نتیجه گرفت که هرچه موقعیت مکانی به مرکز شهر نزدیکتر باشد پس قیمت گرانتری هم دارد. نمودار زیر و خط رگرسیونی بیان‌گر این مفهوم است:

### نمودار همبستگی فاصله خانه از مرکز شهر و قیمت خانه‌ها



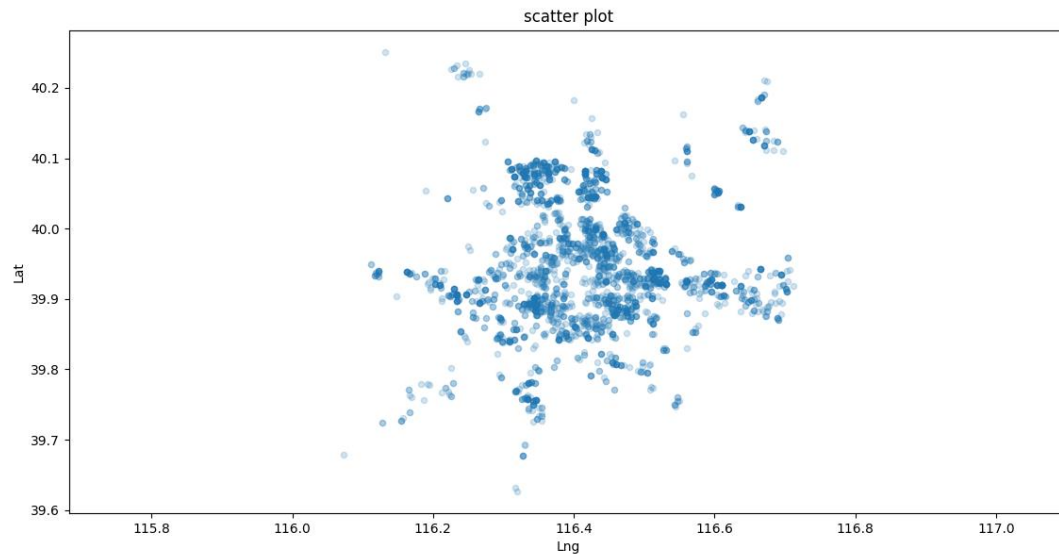
اکنون قصد داریم داده‌ها را به شکل دیگری تحلیل کنیم. میزان اثرگذاری داشتن یا نداشتن آسانسور روی قیمت خانه اهمیت دارد و می‌توان تصمیماتی بر مبنای آن گرفت. نمودار چگالی از مقایسه این دو ویژگی به این صورت می‌باشد:

### تأثیر آسانسور روی قیمت خانه‌ها



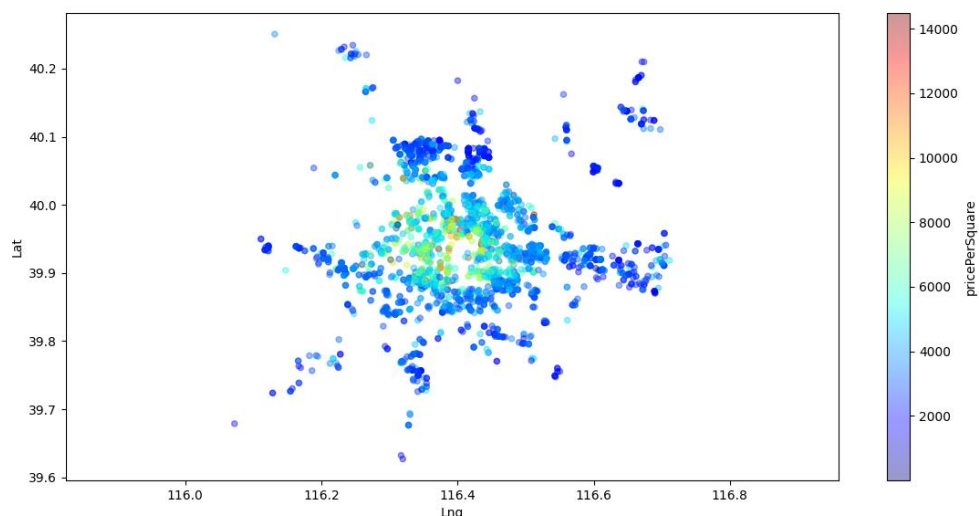
اکنون با تشکر به داشتن مختصات جغرافیایی هر خانه می توان خانه ها را بر اساس موقعیتشان روی نقشه رسم کرد. تراکم و پراکندگی و ویژگی های مختلف می توانند اطلاعات معناداری تولید کنند. نمودار scatter داده ها به شکل زیر است:

### نمودار پراکندگی داده ها



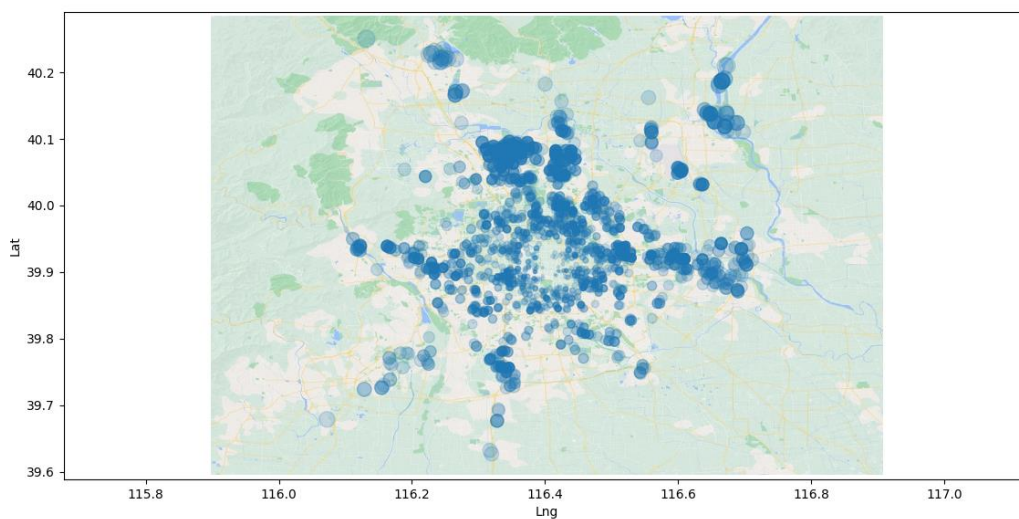
اکنون با استفاده از تعیین رنگ هر نقطه قیمت خانه را مشخص می کنیم. به طوری که هرچی قیمت خانه بیشتر شود، رنگ آن نقطه به سمت رنگ های گرم تر متمایل می شود و هرچه قیمت خانه کمتر شود به سمت رنگ های سردتر متمایل می شود.

## نمودار رنگی شده براساس تغییرات قیمت



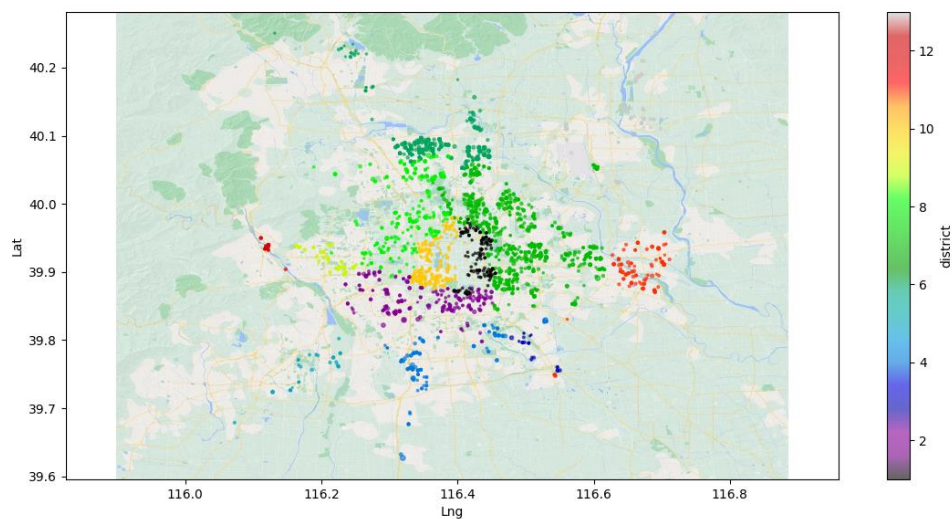
واضح است که هرچه خانه‌ها به مرکز شهر نزدیکتر می‌شوند بر قیمت آن‌ها اضافه می‌شود و هرچه خانه‌ها به حاشیه شهر نزدیکتر می‌شوند از قیمت آن‌ها کاسته می‌شود. با توجه به واقعی بودن موقعیت‌های مکانی می‌توانیم تصویر جغرافیایی شهر پکن را زیر نمودار تصویر کنیم. همچنین اندازه نقاط را معیاری برای ابعاد هر نقطه انتخاب 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:

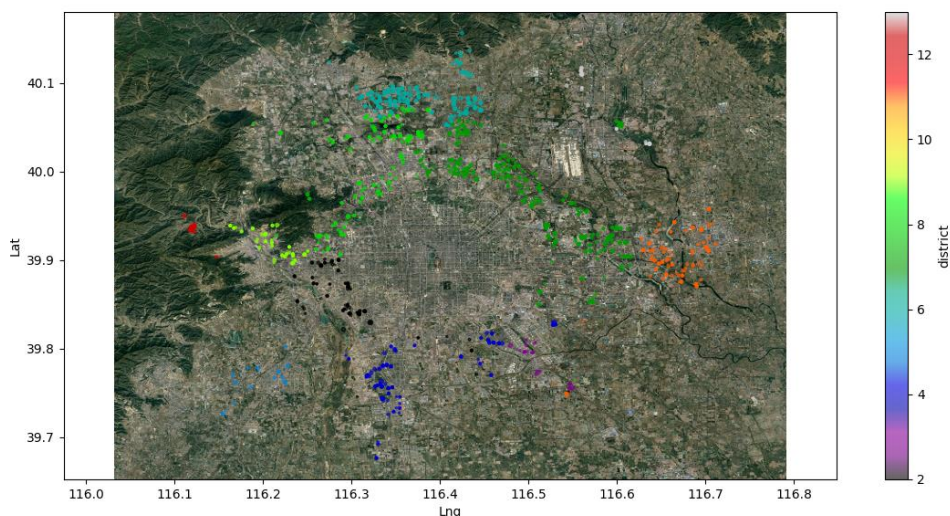
### نمودار داده‌ها بر اساس قیمت و موقعیت مکانی



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.