تحلیل داده های مسکن پلتفرم دیوار

در دنیای پرنوسان بازار املاک، جایی که هر تصمیم می‌تواند سرنوشت‌ساز باشد، داده‌ها چراغ راهی هستند که مسیر را روشن می‌کنند. تحلیل بازار ملک، فراتر از اعداد و ارقام خام، هنر تفسیر داده‌ها و تبدیل آن‌ها به بینش‌های عملی است. داده ها ما را در عمق پیچیدگی بازار مسکن می برد جایی که این سفر را با اتخاذ تصمیمات حرفه ای و سودآور به پایان می رسد. تحلیل بازار ملک از داده های پلتفرم دیوار از (قیمت، مکان، روند) تا مدل سازی و پیش بینی و استراتژی.

1- معرفی داده

در واقع داده‌هات ترکیبی از چند نوع آگهی دیوار هست — فروش، اجازه و اقامت کوتاه‌مدت (مثل ویلا/سوئیت) که به اینصورت می توان دسته بندی کرد.

توضیح	مثال ستون‌ها	دسته
برای تحلیل مکانی و نقشه‌ها	city_slug, neighborhood_slug, location_latitude, location_longitude	موقعیت مکانی
قیمت‌ها و نسبت‌ها	price_value, credit_value, rent_value, total_rent, price_per_meter, rent_price_per_meter	قیمت و اجاره
برای تحلیل نوع ملک	building_size, land_size, rooms_count, floor, total_floors_count, construction_year, has_balcony, has_elevator, has_parking و...	ویژگی‌های فیزیکی
تعیین‌کننده نوع آگهی	cat2_slug, cat3_slug, price_mode, rent_mode, deed_type, has_business_deed	نوع معامله و سند
شخصی یا بنگاه	user_type	نوع کاربر
بررسی روند زمانی	created_at_month	زمان
برای تفکیک لوکس و عادی	has_gas, has_electricity, has_cooling_system, has_heating_system, has_security_guard, has_pool, has_jacuzzi, ...	امکانات و خدمات
مربوط به اقامتگاه‌ها (ویلا، آپارتمان اجاره‌ای)	cost_per_extra_person, rent_price_on_special_days, rent_price_at_weekends, regular_person_capacity, rent_price_on_regular_days,	اجاره کوتاه‌مدت

2- بررسی کیفیت داده

با اجرای کد تعداد و درصد داده های گمشده در هر ستون محاسبه شده اند. و دیتا براساس درصد داده های گمشده از بالا به پایین به صورت صعودی مرتب شده است. همانطوری که در جدول زیر ملاحظه میشود تقریبا تمام یا بیش از 97 درصد داده های مربوط به (rent_to_single, cost_per_extra_person, ... cost_per_extra_person) داده های گمشده هستند که این اطلاعات بیشتر مربوط به اقامتگاه های بومگردی میباشد. بخش دیگری از داده ها مرتبط با دسته رهن و اجاره و تبدیل قیمت ها و اطلاعات دسته سوم که بیشتر حذف شده است شامل خدمات لوکس و لاکچری ساختمان ها شامل استخر و جکوزی و ... میباشد.

columns with highest number of NA's

	column	missing_nums	percent_missing
46	has_pool	970610	97.0610
47	has_jacuzzi	971272	97.1272
48	has_sauna	971521	97.1521
50	property_type	972943	97.2943
52	extra_person_capacity	975991	97.5991
54	rent_price_on_regular_days	981932	98.1932
56	rent_price_at_weekends	986449	98.6449
55	rent_price_on_special_days	989537	98.9537
53	cost_per_extra_person	989759	98.9759
10	rent_to_single	999981	99.9981

و 10 ویژگی که مقادیر بیشتری دارند یا به عبارتی داده های گمشده کمتری دارند در زیر لیست شده اند. به عبارتی می توان گفت که از شاخص های کلیدی برای این بیزینس و قیمت گذاری ملک محسوب می شوند.

columns with lowest number of NA's

	column	missing_nums	percent_missing
0	cat2_slug	0	0.0000
4	created_at_month	0	0.0000
6	description	0	0.0000
1	cat3_slug	1	0.0001
2	city_slug	2	0.0002
7	title	54	0.0054
23	building_size	19606	1.9606
27	rooms_count	154101	15.4101
34	construction_year	184172	18.4172
33	has_parking	271844	27.1844

و براساس تحلیل داده های موجود هیچ دو ردیف تکراری وجود ندارند و این وضعیت مطلوب است به این معنی که یک آگهی با یک مشخصات دوبار ثبت نشده است و این موضوع به صحت شمارش داده ها کمک می کند.

1-2: پیش پردازش داده ها:

قبل از پیش پردازش به تحلیل برخی ناهنجاری های مقداری در زیر پرداخته شده است.

تحلیل ناهنجاری ها و ناسازگاری های مقداری

ناسازگاری های متعددی در نحوه ذخیره سازی مقادیر در ستون ها مشاهده شده است که نیازمند پاکسازی (cleaning) است

1. عدم یکنواختی در متغیرهای دودویی : در ستون هایی که انتظار داریم مقادیر (True/ False) داشته باشیم مقادیر T,f,t,F قابل مشاهده است. این نشان دهنده عدم رعایت حروف کوچک و بزرگ در ثبت داده است و تمام این مقادیر به فرم استاندارد 0و1 یا True , False تبدیل می شوند.
2. ناسازگاری در نوع داده : ستون هایی مانند سال ساخت ساختمان (construction_year) که باید شامل مقادیر عددی صحیح باشند، در برخی موارد شامل رشته عددی است. همچنین در برخی ستون هایی که انتظار عدد می رود متن (Text) ثبت شده و بالعکس.
3. تناقضات منطقی و تاثیر آن بر تحلیل قیمت:

مهم ترین ناسازگاری مشاهده شده، مربوط به تاثیر آن ناسازگاری بر متغیرهای کمی و مدل سازی است. که با محاسبه انحراف چولگی و مقادیر میانگین و میانه مشاهده شده است که حجمی از داده ها در قیمت های بسیار بالا وجود دارند و بعد از بررسی های انجام شده متوجه خطای ثبت داده شده ایم (مثال بارز یک خانه 250 متری در بخش توصیفات ذکر شده، اما در ستون اصلی متراژ واحد، مقدار نادرست 1 متر ثبت شده است. که این تناقض به شدت بر تحلیل مرتبط با قیمت ملک اثر می گذارد زیر اگر متراژ واقعی 250 متر باشد اما در داده ها 1 متر ثبت شده باشد، قیمت اجاره بر متر مربع به طور کاذب 250 برابر بیشتر محاسبه میشود. یا مواردی که سه واحد یک ساختمان ویلایی به فروش می رفته ولی اطلاعات یک واحد ثبت شده است.

4. ناهنجاری عددی و ثبت مقادیر بی معنی: مقادیر غیرواقعی و بسیار بزرگ ثبت شده اند به ازای هر متر که پس از بررسی مشخص شد برخی کاربران صرفا با هدف ثبت یک مقدار پرت یا تست ارقام تکراری 111 را در کنار هم تکرار کرده اند. این ارقام مستقیما به عنوان داده های پرت بسیار شدید عمل می کنند و توزیع قیمت ها را به شدت مخدوش می کند و در مرحله پاکسازی حذف یا اصلاح می شوند.
5. اختلاف قیمت غیر منطقی بین شهرها: در تحلیل میانگین قیمت ها مشخص شد که برخی از شهرهای ناشناخته یا کوچک، دارای میانگین قیمت اجاره ای بسیار بیشتر از شهرهای اصلی و پرتقاضای بزرگ مانند تهران، اصفهان، کرج، مشهد و شیراز دارند، بدون اینکه آپشن یا توجیه لوکس خاصی داشته باشند. که بعد از تمیز کردن داده ها این تناقض تا حد قابل توجهی حل شده است

6. اطلاعات مکانی مفقود: برخی داده ها فاقد لوکیشن و یا شهر و منطقه هستند و این داده ها عملا برای تحلیل های محلی قابل استفاده نیستند.
7. منبع ثبت آگهی : معمولا داده هایی که توسط اشخاص حقیقی ثبت شده اند، دارای داده های پرت بیشتری نسبت به داده هایی که توسط بنگاه ها و املاکی های ثبت شده اند.

**** برای ادامه پروژه نیاز داریم که یک پیش پردازش بر روی دیتافریم صورت پذیرد که شامل استخراج متراژ از متن توضیحات آگهی، حذف ستون های با داده های گمشده زیاد، حذف outliers با استفاده از قیمت و متراژ (که برای این کار از دامنه چارکی استفاده شده است)، استاندارد سازی ستون ها (شامل تبدیل مقادیر فارسی به معادل عددی، جایگزاری سالی متناسب)، تبدیل برخی مقادیر به مقادیر بولی، جایگذاری مقادیر خالی با NaN تا در مراحل بعدی به درستی مدیریت شوند، افزودن ستون متراژ برای هر اتاق، ترکیب حالت های قیمت و اجاره، حذف مقادیر گمشده در ستون هدف، تبدیل مقادیر رهن و اجاره به قیمت واحد(براساس نسبت های بازار مسکن)**

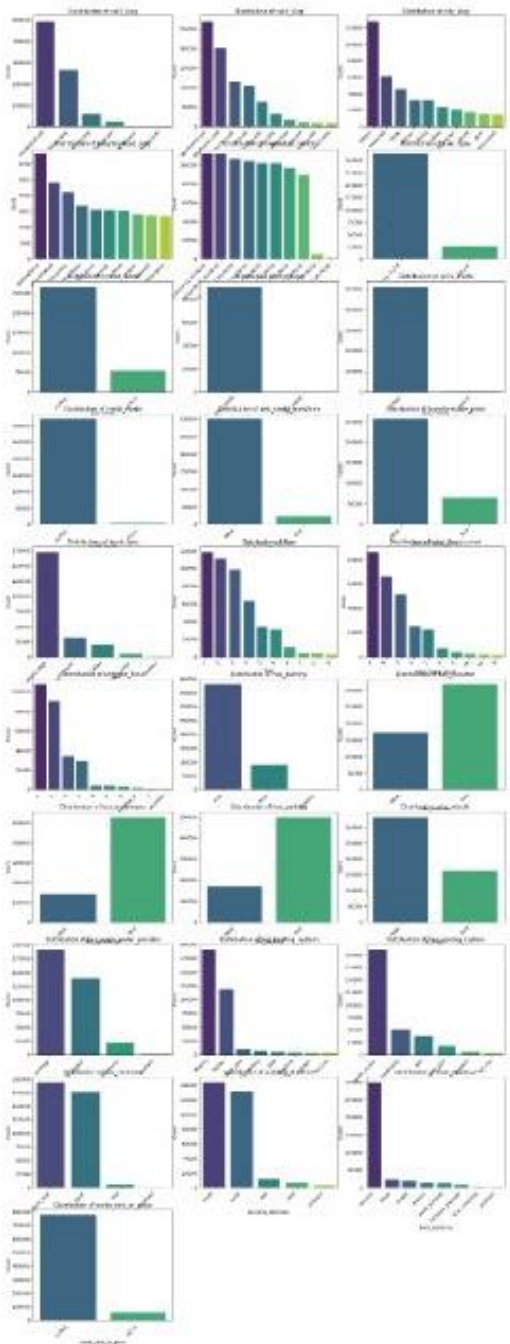
3- تحلیل اکتشافی داده ها (EDA)

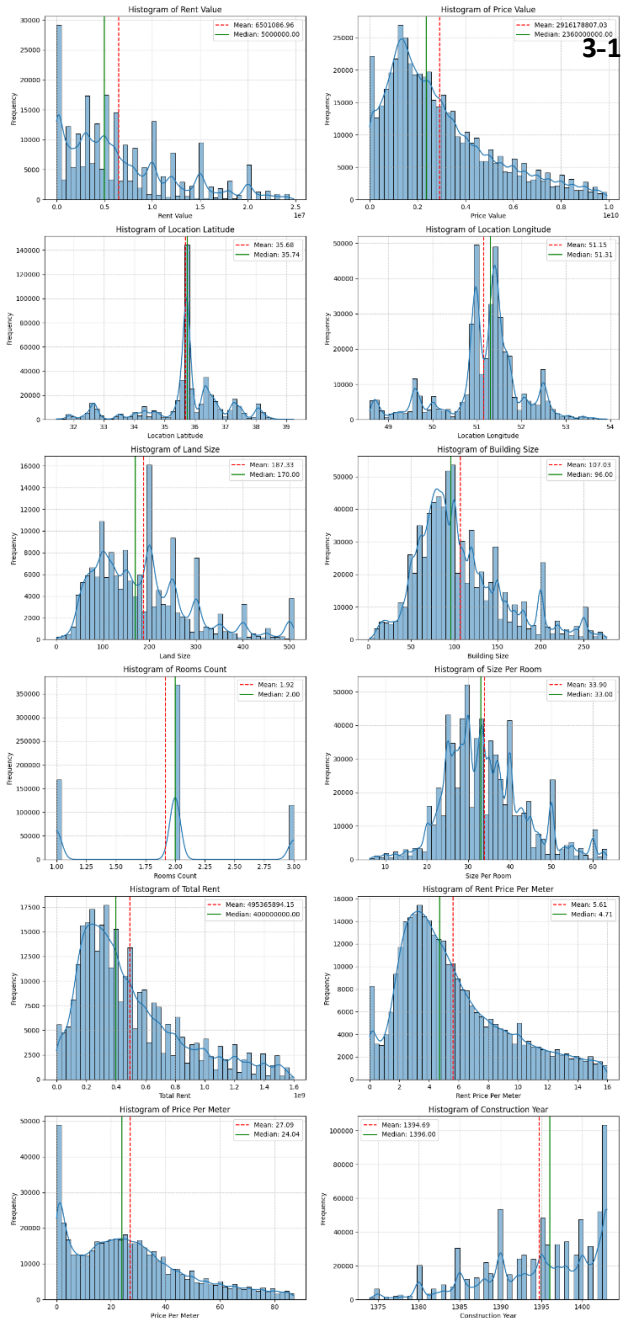
سوال یک: در زیر توزیع و مشخصه‌های آماری هر یک از ستون‌ها (متغیرهای عددی) را ترسیم و بررسی کرده ایم.

1-1 در این بخش با بررسی محتوی ستون ها ابتدا تعیین کردیم که هر ستون متعلق به کدام دسته (عددی، دسته بندی، بولی و متنی) می باشد (دلیل اهمیت تفکیک ستون ها به این دلیل است که این فرآیند برای تمیزکاری و مدلسازی ضروری است، مثلا ستون های عددی با میانگین پر می شوند، در حالی که ستون های دسته بندی برای تبدیل شدن به فرمت قابل درک برای مدل های یادگیری ماشین آماده می شوند).

2-1 مشخصه های آماری ستون های عددی در زیر آمده است. این مشخصه ها شامل توصیف آماری تعداد،میانگین،انحراف معیار،کمترین، بیشترین و چهارک ها می باشد. علاوه بر آن چولگی(عدم تقارن توزیع داده ها) و کشیده گی نیز آورده شده است. چولگی و کشیدگی به ما کمک می کنند که متوجه شویم آیا ستون های اصلی (مانند قیمت/ متراژ) نیاز به تبدیل لگاریتمی دارند یا خیر تا توزیع برای استفاده در مدل های خطی (مانند رگرسیون) به حالت نرمال تر نزدیکتر شود.

همچنین نمودار توزیع فراوانی و ستون های دسته بندی (نمایش تعداد آگهی ها در هر دسته یا کلاس(فقط 10 دسته پرتکرار اول)) در زیر نمایش داده شده است در خصوص اینکه کدام دسته بیشترین تعداد آگهی را به خود اختصاص داده است.





تغییرات زیاد در فراوانی دسته ها به این معنی است که داده ها حول یک دسته خاص (مانند نوع ملک خاص، منطقه خاص، حالت قیمت گذاری غالب) متمرکز شده است. بعضی از نمودارها که فقط دو میله دارند در ستون های امکانات ارزش های بولین (دارد، ندارد) را نشان می دهند.

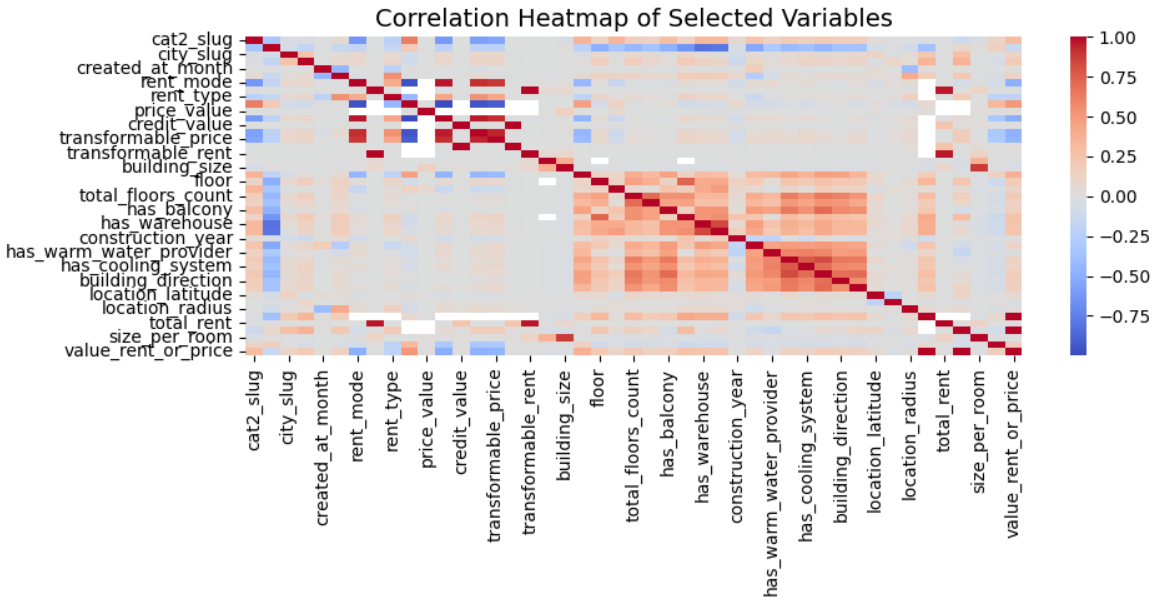
همانطور که در بخش چولگی مشاهده می شود این داده ها در برخی نمودارها، بسیار چوله هستند و مقادیر پرت زیادی دارند. چولگی میزان عدم تقارن توزیع نشان میدهد که در اینجا، چولگی تقریبا در تمام ستون های مالی و متراژ بسیار شدید و مثبت است. مقادیر بالای کشیدگی (به خصوص در ستون های مالی و اندازه ها) نشان دهنده شکست فرضیه نرمال بودن و وجود مقادیر پرت افراطی است که می تواند مدل های یادگیری ماشین را منحرف کند.

این نمودار چولگی توزیع داده های قیمت بر متر مربع است که فراوانی قیمت های مختلف را نشان می دهد. چولگی به سمت راست: توزیع به شدت به سمت راست کشیده شده است و این احتمالا به این معناست که بیشتر داده ها در سمت چپ (مقادیر پایین تر) متمرکز شده اند، اما تعداد کمی از مقادیر بسیار بالا نیز وجود دارند که دنباله توزیع را به سمت راست طولانی کرده اند. همچنین این نمودار نشان می دهد بازار ملک مورد بررسی دارای تفاوت قیمت بسیار زیاد است و اگر هدف مدلسازی باشد توزیع به شدت نرمال نیست و برای بهبود عملکرد مدل های خطی، احتمالا با تبدیل لگاریتمی روی ستون price per Meter اعمال شود تا چولگی کاهش یابد.

1-4 در این بخش با محاسبه IQR دامنه بین چارکی را محاسبه کرده ایم و داده ها را طوری فیلتر می کنیم که فقط مقادیری باقی بمانند که بین مرز پایین و بالا قرار دارند. (این کار مقادیر پرت شدید را حذف می کند، همان ناهنجاری هایی که باعث چولگی و کشیدگی بالا شده بودند)

سوال دو:

1-2: برای بررسی ارتباط دودویی تعدادی از متغیرها و کشف همبستگی های مهم بهترین ابزار در اینجا نقشه حرارتی (Heatmap Correlation) است. نقشه حرارتی همبستگی روابط خطی بین متغیرها را نشان می دهد. در این نمودار جفت ستون هایی با همبستگی بیش از 0.8 نمایش داده شده است. این نقشه حرارتی همبستگی متغیرها نشان می دهد که تعدادی از متغیرها ارتباط معنایی مثبت و تعدادی دیگر ارتباط معنایی منفی دارند.



نمودار جفت هایی با کوریلیشن بیش از 0.8

2-1-1: دسته اول: دو تایی هایی با همبستگی بسیار قوی (قرمز تیره)

* زیر جفت هایی که همبستگی بسیار بالایی دارند و همبستگی 1 دارند لیست شده است.

(transformed_credit , credit_value) و (transformable_credit - credit_value) و (transformable_rent - transformed_credit) و این شامل متغیرهای مالی که قابل تبدیل به یکدیگر هستند می باشد. و وجود همبستگی یک به این معناست که اطلاعات یکسانی را به مدل می دهند.

** رنگ قرمز تیره همبستگی مثبت قوی که با افزایش یک متغیر دیگری افزایش میابد: به عنوان مثال (ransformed_rent, rent_value) و (rent_price_at_weekends, rent_price_at_weekends) و یا شاخص های امکانات رفاهی تقریبا در وسط تصویر (has_balcony, has_parking, has_sauna, has_barbecue, has_pool, has_restroom) همبستگی مثبت زیادی با یکدیگر نشان می دهند یا به عبارتی ارزش ملک و قیمت اجاره آن را افزایش می دهد که به عنوان **همبستگی خوشه امکانات Amenity Cluster) (Correlation) شناخته می شود. به عبارت دیگر خانه هایی که امکانات لوکس بیشتری مانند (سونا، استخر، باربیکیو) دارند معمولا تمایل دارند که همه ویژگی ها را با هم داشته باشند.**

2-1-2: دسته دوم: جفت های دارای همبستگی معکوس بسیار قوی (آبی تیره)

رنگ آبی تیره همبستگی منفی قوی که با افزایش یک متغیر دیگری کاهش می یابد:

* rent_mode و credit_mode (که تقریبا یکی هستند) به شدت با price_value رابطه معکوس دارند. به این معنی که اگر یک ملک در حالت اجاره قرار داشته باشد، قیمت فروش آن باید صفر باشد، بنابراین با افزایش حالت اجاره، قیمت فروش کاهش می یابد.

متغیرهایی که بیشتر در سمت چپ و بالای تصویر دیده می شود، شامل **(rent_credit_transform, value)** همبستگی منفی قابل توجهی نشان می دهند و ممکن است به این معنی باشد که املاک با حالت اجاره تمایل به قیمت گذاری پایین تری دارند.

3-1-2: دسته سوم: جفت های دارای همبستگی ضعیف یا عدم همبستگی:

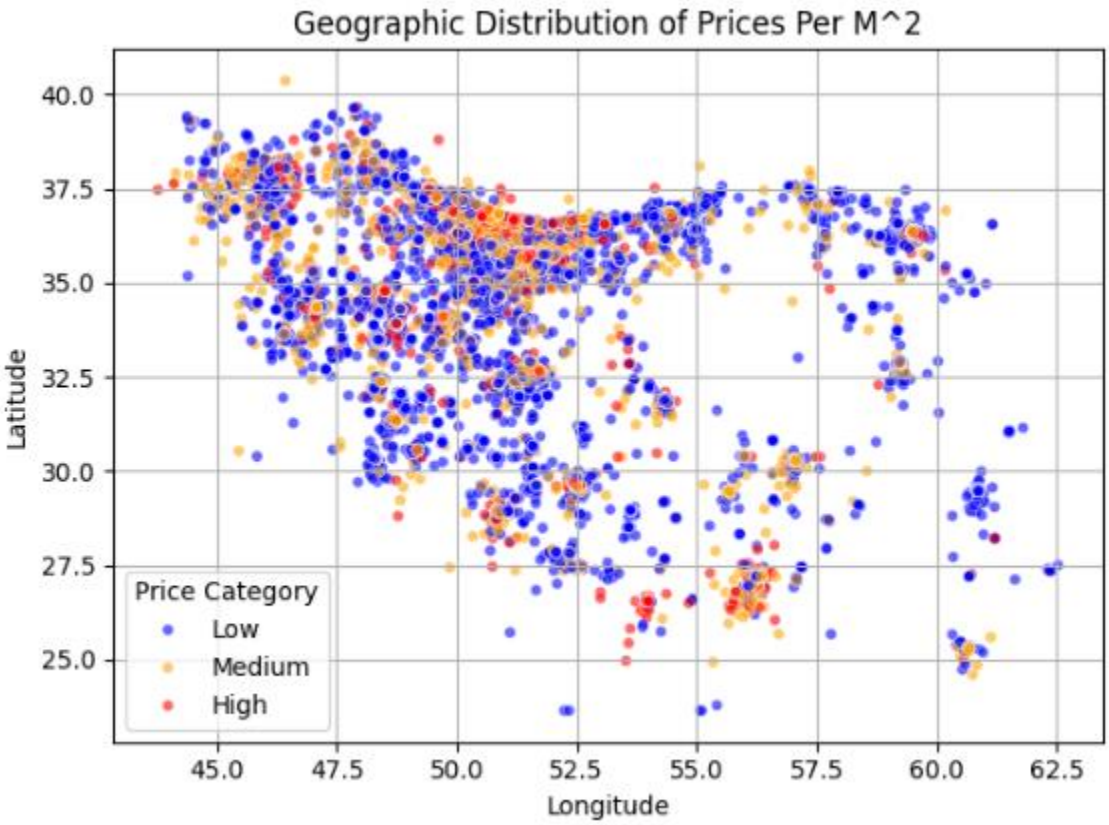
بیشتر ویژگی ها با این مشخصه توزیع شده اند که یا به علت ارتباط کم سایر متغیرهاست و یا به این علت است که متغیرها همبستگی غیرخطی دارند.

سوال سه:

نکات قابل توجهی را که از بررسی توزیع‌ها و نمودارهای پراکندگی و مقدار همبستگی دریافتید فهرست شده است. هدف این بخش درک توزیع قیمت ها براساس موقعیت جغزافیایی و مقایسه قیمت ها در شهرهای ایران است.

* در این بخش ردیف هایی که مختصات جغرافیایی آن ها صفر بوده حف شده اند.

* داده ها برای فروش آپارتمان و مسکونی فیلتر شده اند



با توجه به نمودار و آگهی های ارائه شده در پلتفرم دیوار ، این نمودار پراکندگی توزیع جغرافیایی قیمت های اجاره برحسب مترمربع در سه دسته پایین- آبی، متوسط- نارنجی، و بالا- قرمز را نشان می دهد که محور افقی طول جغرافیایی و محور عمودی عرض جغرافیایی را نشان می دهد

بیشتر تراکم ها در شمال نقشه، تقریبا در عرض جغرافیایی بالاتر از 30 و طول جغرافیایی بین 45 تا 55 درجه مشاهده می شود که احتمالا مربوط به مناطق پرجمعیت و متمرکز شهری (شامل پایتخت و شهرهای بزرگ مرکزی) است که تعداد بیشتری آگهی ثبت کرده اند.

مناطق کم تراکم: نقاط پراکندگی کمتر در سمت راست نقشه بیشتر دیده میشود با طول جغرافیایی بیشتر و عرض جغرافیایی حدود 25 درجه که نشان دهنده فعالیت کمتر می باشد.

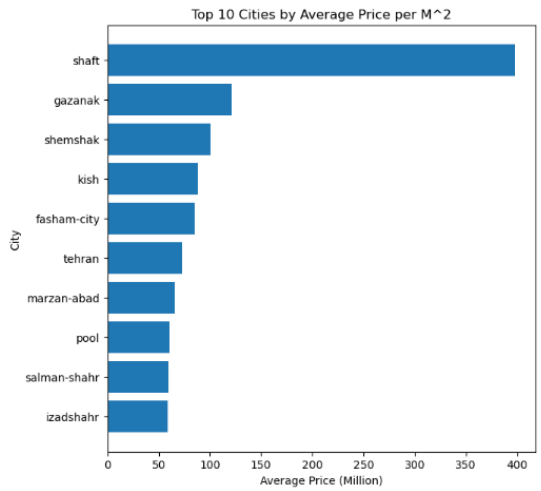
در دسته بندی قرمز که الگوهای قیمت گذاری گران قیمت را مشخص می کند تمرکز اصلی در شمال نقشه و به ویژه در مناطق پرتراکم است و این نشان می دهد که گران ترین املاک در شهرهای پرجمعیت و توسعه یافته شمالی قرار دارند. و دسته دیگری در نواحی ساحلی جنوبی قابل

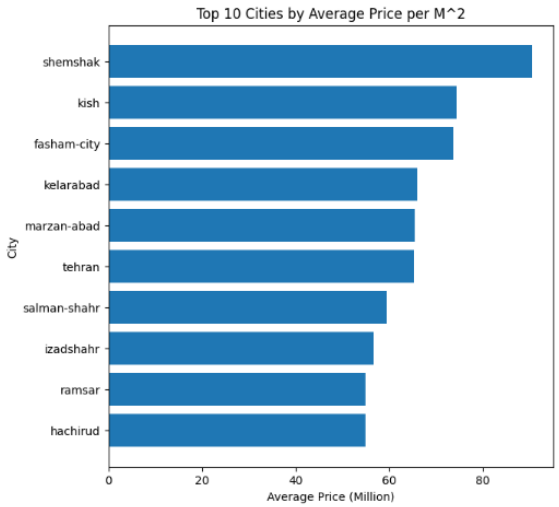
مشاهده است که ممکن است شامل جزایر یا شهرهای ساحلی خلیج فارس باشد و به دلیل جذابیت توریستی یا خاص بودن منطقه قیمت بالایی دارند. این نمودار نشان می دهد ارزش ملک با موقعیت جغرافیایی ارتباط نزدیکی دارد.

تحلیل قیمت براساس شهرها

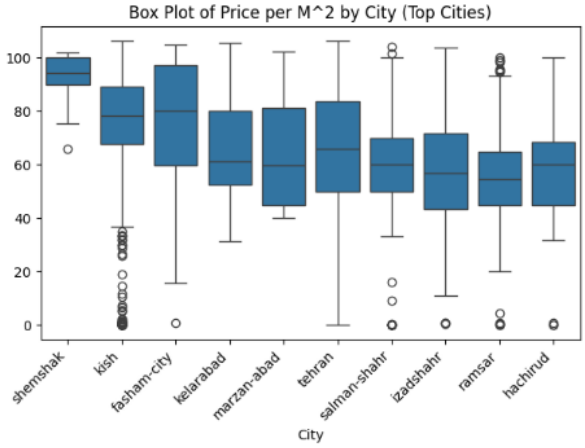
داده خام به‌شدت آلوده به مقادیر پرت بوده و نمی‌توان به آن تکیه کرد و براساس داده های تمیز شده می توان تصمیم سازی نمود. برای مقایسه اهمیت پاکسازی داده ها برخی موارد قبل و بعد از پاکسازی با یکدیگر مقایسه شده اند.

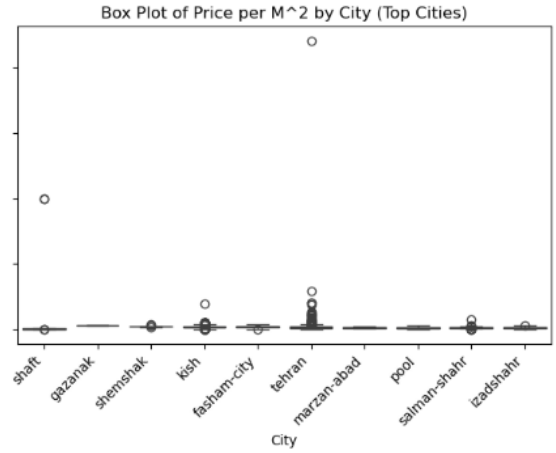
نمودار میانگین قیمت فروش هر متر در ده شهر گرانتر : این نمودار نشان می دهد که 10 شهر با بیشتری میانگین قیمت هر متر مربع کدام هستند و محور از بالا به پایین مرتب شده (گرانترین در بالا قرار دارد)





تحلیل نمودار قبل و بعد از پاکسازی داده ها	
قبل از پاکسازی داده ها	بعد از پاکسازی داده ها
در نمودار سمت راست (قبل از پاکسازی)، برخی شهرها قیمت‌های غیرعادی و بسیار بالایی دارند که باعث شده میانگین قیمت آن‌ها به شدت بالا برود. شهر **shaft**با میانگین حدود **400میلیون تومان** برای هر متر مربع در صدر قرار گرفته است؛ این مقدار کاملاً غیرطبیعی است. شهر **gazanak**با اختلاف زیاد و حدود **150میلیون تومان** در جایگاه دوم است. سایر شهرها مانند **shemshek**، **kish**و **tehran** در محدوده قیمتی 40–70 میلیون قرار دارند. فاصله بسیار زیاد shaft از بقیه شهرها نشانگر وجود **چند رکورد پرت (Extreme Outliers)** است. چنین قیمت‌هایی نه با بازار واقعی همخوانی دارند و نه با سایر دیتاهای موجود. این Outlier ها باعث شده‌اند: میانگین شهرها گمراه‌کننده شود، تحلیل مکان‌محور غلط شود، رتبه‌بندی شهرها فاقد اعتبار شود.	در نمودار سمت چپ (بعد از پاکسازی)، قیمت‌ها در محدوده‌ای منطقی قرار گرفته‌اند و رتبه‌بندی شهرها واقعی‌تر شده است. صدر جدول اکنون شهر **shemshek**با حدود **90میلیون** است؛ عددی منطقی و همخوان با بازار مناطق کوهستانی لوکس می باشد. شهرهای **kish**، **fasham-city**، **marzan-abad**در رتبه‌های بعدی قرار دارند. تهران نیز در نزدیکی این محدوده قرار دارد و اختلاف آن با سایر شهرها منطقی است.





نمودار Box plot: این نمودار پراکندگی قیمت، میانه، بیشترین و کمترین قیمت هر متر در 10 شهر گران کشور را نشان می دهد و با استفاده از این نمودار می توان فهمید در هر شهر قیمت چقدر نوسان دارد و اینکه آیا شهرهایی با اختلاف شدید قیمت وجود دارد یا خیر و میانه قیمت هر متر در هر شهر کجاست.

برای توضیح بیشتر شهرهایی مانند **شِمشک، کیش و فشم** بالاترین قیمت هر متر مربع را دارند و تمرکز قیمت در آن‌ها در سطوح بالا و با پراکندگی کمتر قرار گرفته است. که این سه منطقه جزو مناطق لوکس به حساب می آیند. البته در منطقه کیش تعدادی ملک با قیمت متوسط هم هست که در صورت نیاز می توان بررسی تفکیکی قیمت براساس «موقعیت نزدیک به ساحل و سال ساخت و ... را بررسی نمود.

در شهر هایی مثل فشم و تهران پراکندگی چشمگیر است که ممکن است به تنوع زیاد در کیفیت و نوع ملک یا ناشی از اعلان های اشتباه املاک در برخی مناطق باشد.

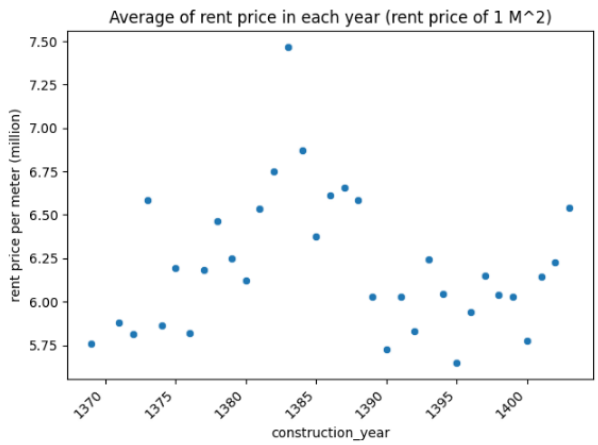
در منطقه رامسر پراکندگی زیاد و آوتلایرهای متعدد ، بازار متنوعی با اختلاف قیمت چشمگیر بین املاک معمولی و لوکس ایجاد کرده است

* خروجی زیر نشان دهنده این است که هر یک از آگهی های فروش یا اجاره مربوط به چه نوع مسکن می شود

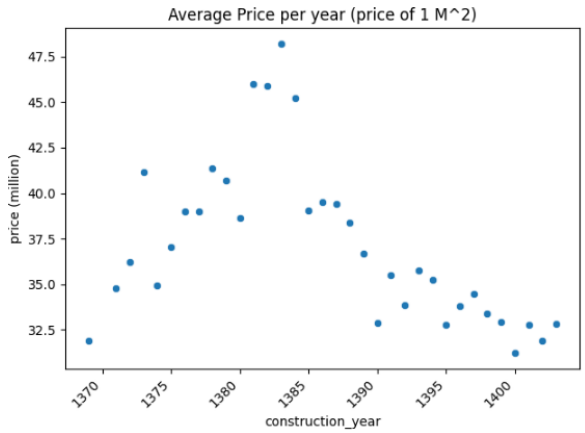
```
cat2_slug            cat3_slug
commercial-rent      shop-rent                          32936
                     office-rent                        17174
                     industry-agriculture-business-rent  8761
commercial-sell      shop-sell                           9836
                     industry-agriculture-business-sell  8932
                     office-sell                         2710
real-estate-services presell                                5
                     partnership                            1
residential-rent     apartment-rent                    200395
                     house-villa-rent                   62853
residential-sell     apartment-sell                    270543
                     house-villa-sell                  116019
                     plot-old                          105735
temporary-rent       villa                                  2
                     suite-apartment                        1
Name: count, dtype: int64
```

* نمودار زیر توضیح دهنده متوسط قیمت فروش یک متر مربع از آگهی های فروش براساس سال ساخت است.

rent_per_meter



price_per_meter-sales

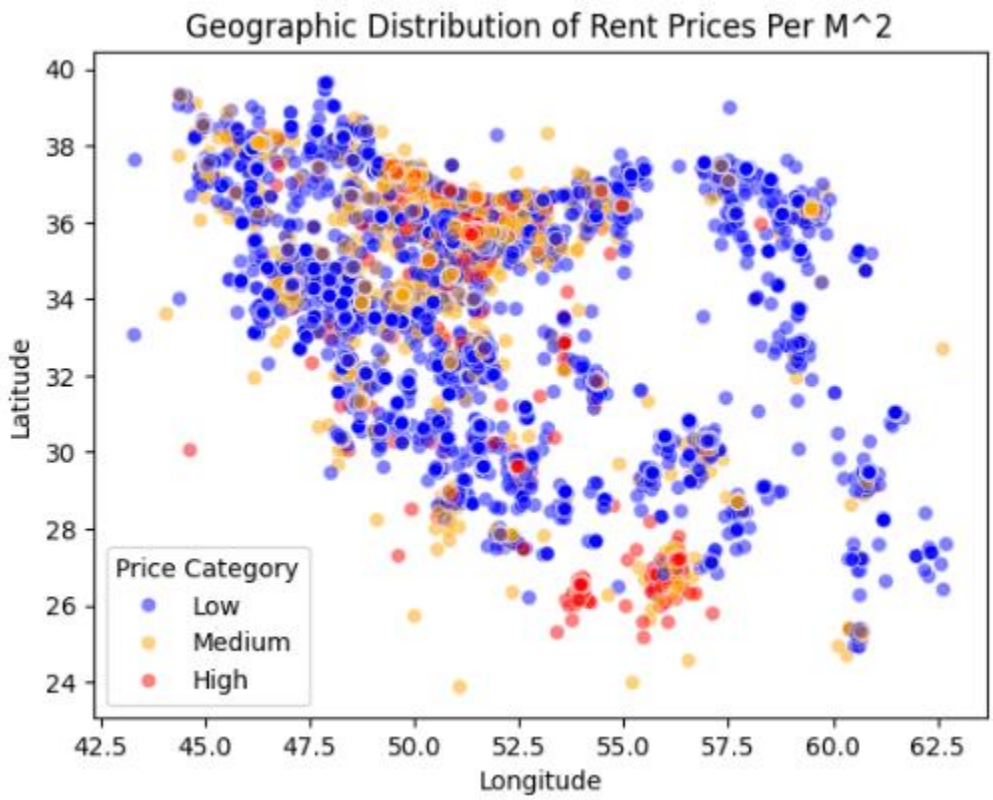


این دو نمودار پراکندگی رابطه بین سال ساخت و میانگین قیمت برای فروش و اجاره برحسب متر مربع را نشان می دهد.

نمودار سمت چپ: سازه های ساخته شده در بین سال‌های **۱۳۷۰ تا ۱۳۸۵**، میانگین قیمت اجاره در این دوره در محدوده باریکی (حدود ۵.۷۵ تا ۶.۶ میلیون) با نوسانات زیاد حرکت می‌کند. اوج قیمت مربوط به سازه های 1384 است که به قیمت نزدیک به ۷.۵ میلیون رخ داده است. سازه های جدیدتر به ازای هر متر قیمت گذاری کمتری شده اند که این ممکن است عوامل متعددی شامل محله و امکانات زیرساختی و ... داشته باشند. و مشاهده می کنیم که ساختمان های جدیدتر لزوما بالاترین قیمت اجاره را ندارند.

نمودار سمت راست: این نمودار رابطه بین سال ساخت و میانگین قیمت فروش بر متر مربع را نشان می‌دهد و مشابه تحلیل قبل بیشترین میانگین قیمت فروش متعلق به ساختمان های جدید نیست و بیشترین قیمت فروش متعلق به سازه هایی که حدود 1385 تا 1390 ساخته شده اند می باشد که دلایل متعددی مانند لوکیشن های کلیدی و مرغوب باشند که ارزش خود را حفظ کرده یا افزایش داده اند یا ناشی از اطلاعات اشتباه باشد که در مرحله اول قید شد از جمله اینکه گاهی متراژ زمینی در ستون توضیحات 250 متر قید شده بود ولی در ستون مربوطه یک متر و این موارد این چنینی موجب ایجاد اوتلایرها می شوند.

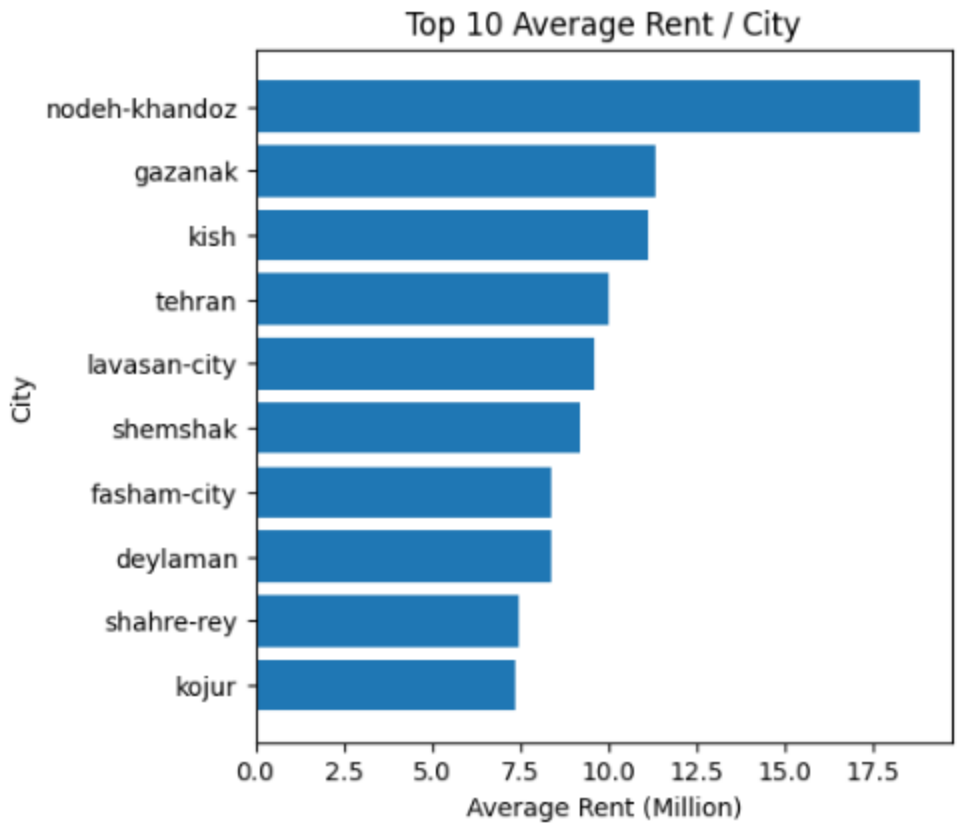
نمودار زیر توضیح دهنده نحوه توزیع قیمت اجاره یک مترمربع از اجاره آپارتمان های مسکونی در نقاط مختلف کشور



تحلیل: نمودار پراکندگی جغرافیایی توزیع قیمت اجاره بر متر مربع را در ایران نشان می دهد. این نمودار به ما کمک می کند تا الگوهای منطقه ای و تمرکز بازارهای اجاره گران قیمت در سراسر کشور را مشاهده کنیم. بیشترین تراکم نقاط در ناحیه شمالی کشور و بیشتر منطبق بر موقیت استان های شمالی، تهران، البرز و شهرهای بزرگ مرکزی می باشد. وجود نقاط قرمز و نارنجی خارج از مناطق متراکم، ممکن است نشان از شهرهای کوچک توریستی یا آگهی های پرت در منطقه باشد. همچنین این نمودار نشان می دهد که در مناطق متراکم، اجاره بها یکنواخت نیست و نقاط آبی و نارنجی و قرمز در هم آمیخته نشان از این است که در مناطق جغرافیایی کوچک مانند یک شهر، آپارتمان اجاره ارزان، متوسط و گران در کنار یکدیگر هستند و ممکن است عوامل محلی تر مانند دسترسی محلی، امکانات خاص و نوع محله) تاثیر زیادی بر قیمت اجاره داشته باشند.

نمودار 10 شهر برتر براساس میانگین اجاره

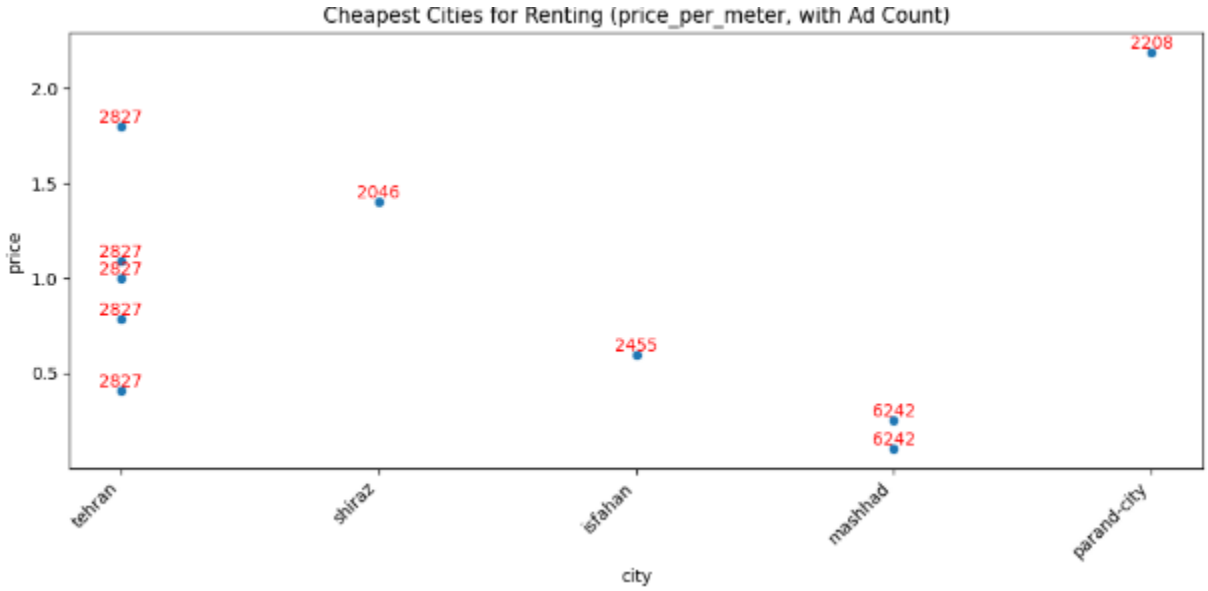
nodeh-khandozبا اختلاف قابل توجهی نسبت به سایر شهرها، بالاترین میانگین اجاره‌بها (نزدیک به ۱۸ میلیون تومان) را دارد. این نشان دهنده این است که این منطقه یا دارای تعداد محدودی آگهی بسیار گران قیمت است که میانگین را بالا برده، یا یک منطقه بسیار لوکس (ویلا، گردشگری) است یا شامل داده های پرت است. کیش به عنوان یک منطقه آزاد و گردشگری رقمی منطقی در بین گران ترین ها دارد.


Top 10 Average Rent / City
nodeh-khandoz
gazanak
kish
tehran
lavasan-city
shemshak
fasham-city
deylaman
shahre-rey
kojur
City
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Average Rent (Million)

4- تحلیل آگهی های املاک و بازار املاک ایران

دینفع 1- مستاجر و خریدار:

در ابتدا پاکسازی و حذف ناهنجاری ها اتفاق افتاده است. (تمام آگهی هایی که قیمت اجاره بر متر مربع آنها صفر یا منفی است حذف شده اند، دامنه بین چارکی IQR برای ستون rent_price_per_meter محاسبه شده و مرزهای استاندارد بالا و پایین برای حذف ناهنجاری ها تعیین شده اند و خلاصه ناهنجاری قیمتی افراطی پاک شده اند. همچنین در این بخش شناسایی ارزان ترین ها (تمام مواردی که قیمت اجاره آن ها کمتر از Q1 است انتخاب شده اند) که مرز قیمت گذاری برای 25 درصد ارزانترین املاک هستند و حجم بازار ارزان در آن شهرها تعیین شده اند.



ارزانترین شهرها برای اجاره مسکن:

این نمودار پراکندگی، میانگین قیمت اجاره بر متر مربع (rent_price_per_meter) را برای تعدادی از شهرها نشان می‌دهد. همچنین، تعداد آگهی‌های موجود برای هر شهر (Number of Ads) به صورت عدد قرمز در بالای هر نقطه نمایش داده شده است.

هدف: شناسایی شهرهایی که کمترین میانگین قیمت اجاره بر متر مربع را دارند. و شناسایی ارزانترین شهر ها و تعداد آگهی ها نشان می دهد در شهرهایی که قیمت بالاتری به ازای هر متر است تعداد آگهی بیشتری ثبت شده است و به علت حجم بالای داده های ثبت شده دارای اعتبار بیشتری هستند و البته شهرهایی با تعداد آگهی بسیار کم اعتبار پایین تری دارند و ممکن است با افزایش آگهی تغییر زیادی در قیمت داشته باشند.

تحلیل کد ذینفع مستاجر: در بخشی از کدها میانگین اجاره براساس شهر محاسبه شده است که تابع mean برحسب مترمربع در هر شهر اعمال شده است و این آمار برای مقایسه کلی هزینه اجاره بین شهرها مفید است. برخی از اطلاعات در زیر قابل مشاهده هاست.

	city_slug	avg_rent_city
0	Iranshahr	1.865820
1	Kordkuy	2.557441
2	aalasht	0.775000
3	abadan	2.857721
4	abadeh	2.526974
...	...	...
402	zargarshahr	4.857143
403	zarghan	2.392437
404	zibakenar	4.474691
405	ziraab	3.116194
406	ziyabar	2.138889

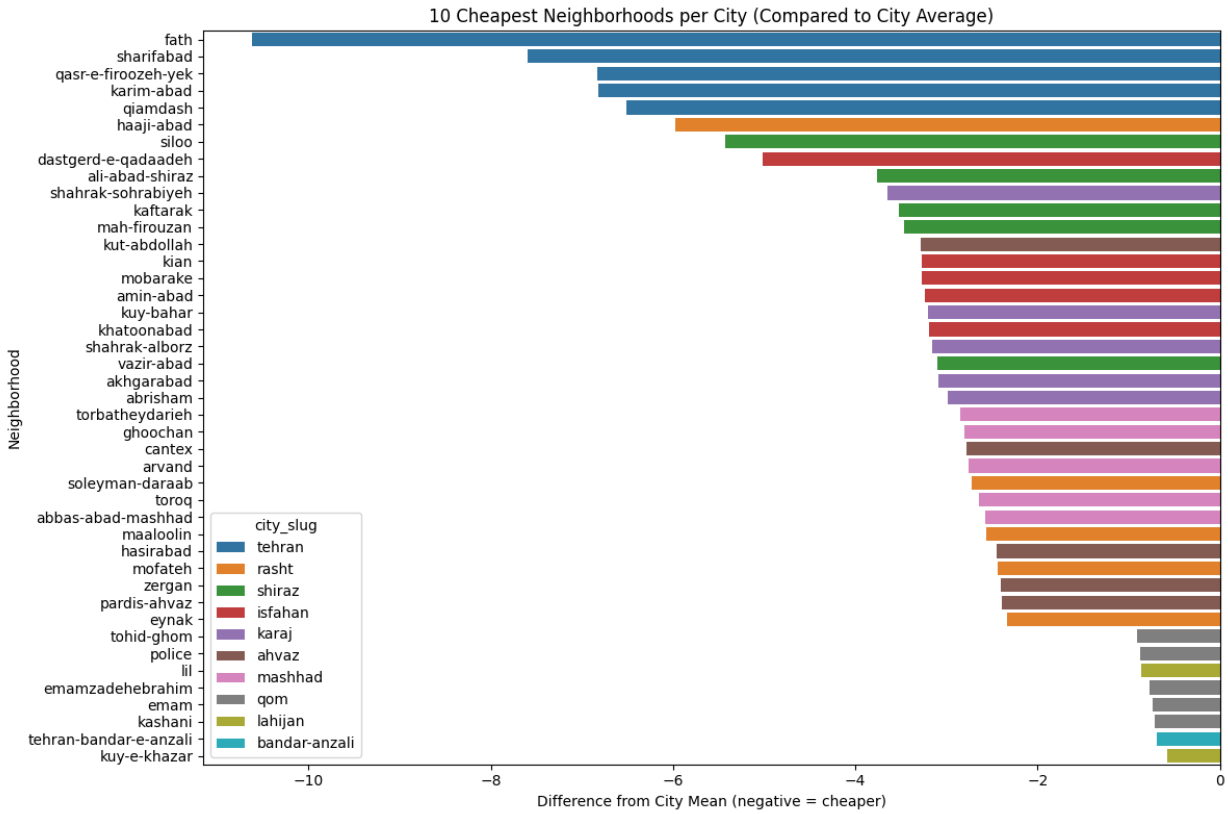
407 rows × 2 columns

شناسایی محله های مقرون به صرفه : در زیر مقایسه مستقیم بین میانگین اجاره یک محله و میانگین شهر متبوع خود را در یک ردیف ممکن می سازد. هدف این است که فقط محله هایی که میانگین قیمت اجاره بر متر مربع آن ها از میانگین کل قیمت اجاره شهر خود کمتر است، نگهداری شوند. تهران با بیش از دوبرابر محله های زیرمیانگین نسبت به شهر دوم بیشترین تنوع محله ای را برای پیدا کردن اجاره ارزانتر فراهم می کند که البته ممکن است به علت این باشد که تعداد آگهی های ثبت شده در پلتفرم دیوار در تهران بیشتر از سایر شهرها است.

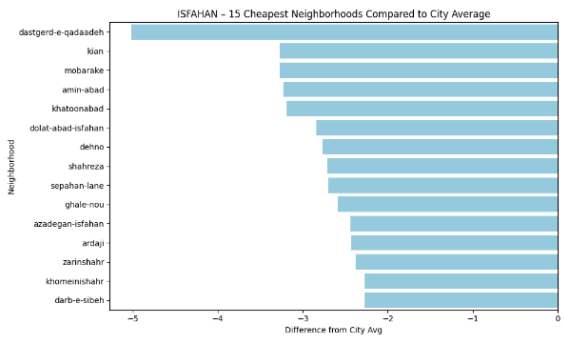
```
city_slug
tehran          221
isfahan         133
shiraz          126
mashhad          96
karaj            56
ahvaz            42
rasht            36
qom              18
lahijan           2
bandar-anzali     1
```

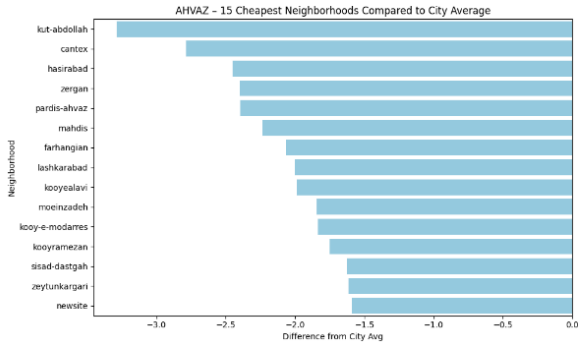
شناسایی ارزانترین محله های شهر :

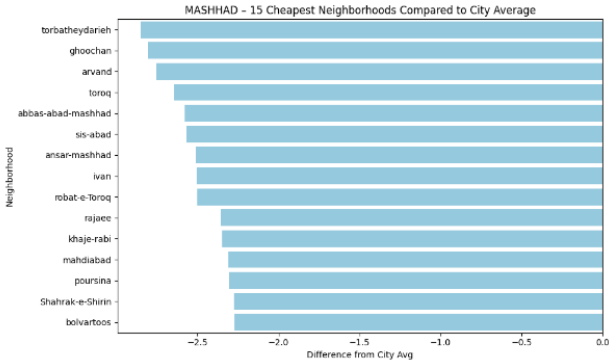
نمودار زیر 5 محله ارزانتر در هر شهر را نشان می دهد که بیشترین اختلاف قیمت منفی (یعنی ارزانترین) را نسبت به میانگین قیمت اجاره کل شهر خود دارند که به خریداران/ مستاجران کمک می کند تا نقاط با بهترین معامله را پیدا کنند. در واقع این نمودار کمک می کند، محله‌هایی را پیدا کنند که **نسبت به سطح عمومی گران‌قیمت شهر خود، به طور نامتناسبی ارزان‌تر** هستند

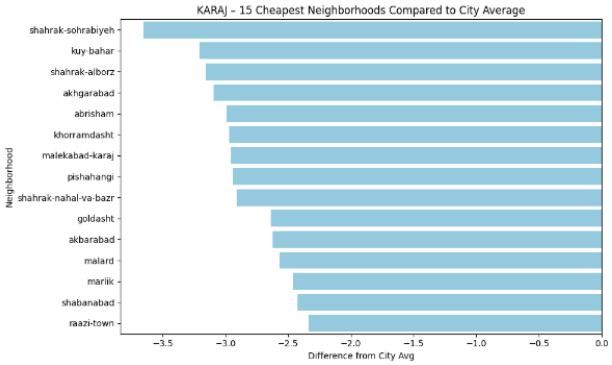


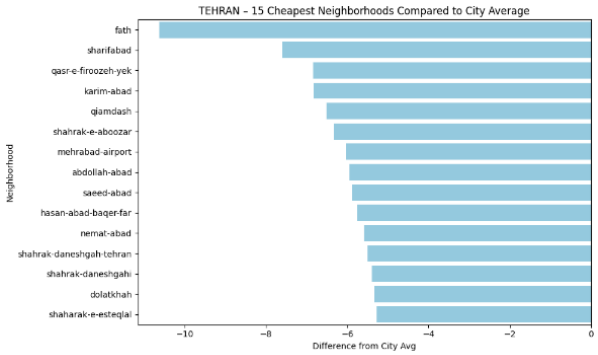
و در زیر چند نمودار تفصیلی و عمیق تر از ارزانترین گزینه های اجاره در هر شهر است که در فایل اصلی تماما لیست شده است.

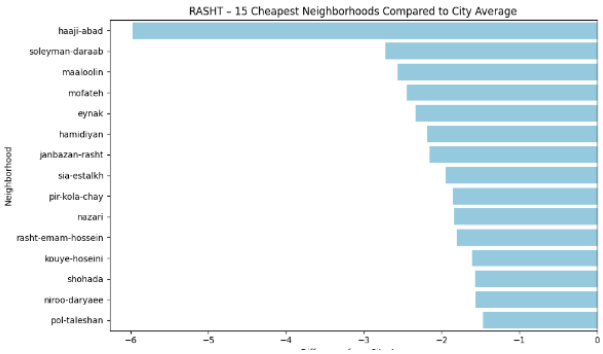












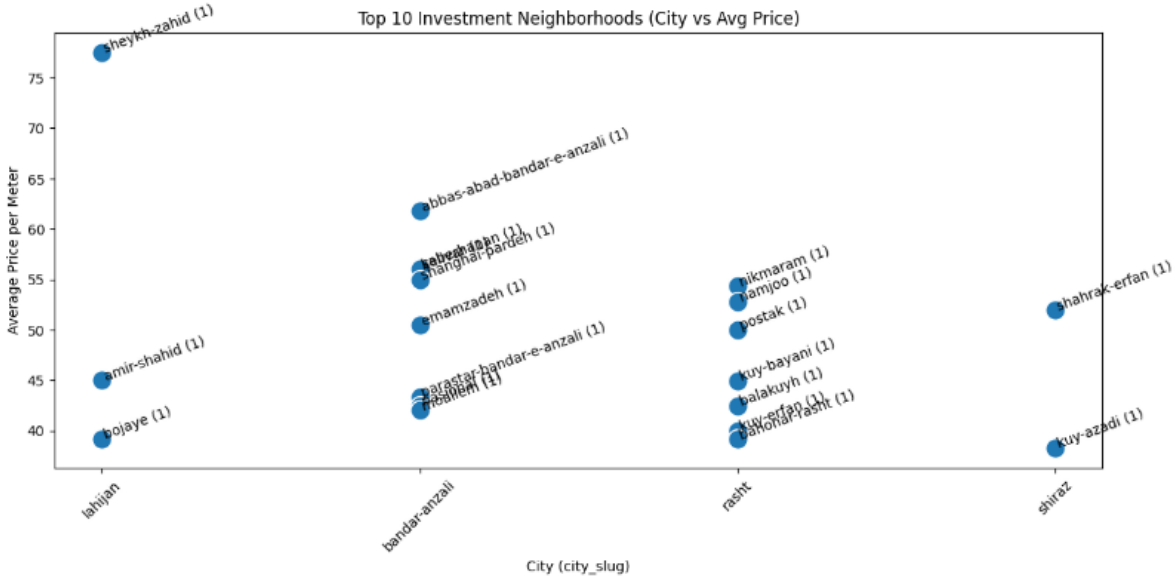
ذینفع2- سرمایه گذار:

شناسایی محله‌های سودآور سرمایه‌گذاری: باید دو شاخص کلیدی را به‌صورت همزمان تحلیل کنیم، بنابراین هدف پیدا کردن محله هایی است که:

- برای شناسایی محله‌های ایده‌آل سرمایه‌گذاری با دو شرط **رشد پایدار قیمت** (این شاخص نشان می‌دهد که آیا سرمایه‌گذاری در یک محله ارزش خود را در طول زمان حفظ کرده و افزایش می‌دهد یا خیر) و **عرضه کم (تقاضای بالا که منجر به کمتر خالی**

ماندن ملک و افزایش پتانسیل افزایش قیمت می‌شود)، داده‌ها را در دو بعد کلیدی تحلیل و سپس محله‌ها را بر اساس این شاخص‌ها رتبه‌بندی کردیم

- و نرخ خالی ماندن ملک در اینجا با استفاده از تعداد آگهی تخمین زده میشود.



10 محله برتر برای سرمایه گذاری در نمودار بالا در چندین شهر مختلف قابل مشاهده است محله های برتر در شهر های لاهیجان، بندرانزلی، رشت و شیراز پراکنده شده اند و تعداد محله های برتر با نام در هر شهر مشخص شده است. لاهیجان محدوده قیمتی بالاترین میانگین قیمت در بین کل محله های برتر را دارد و بندرانزلی دارای بیشترین نوسان قیمت در محله هایش است و شیراز در محدوده قیمتی پایینتری قرار دارد که البته باز تعداد آگهی های کمتر قابل تامل است و نیاز به بررسی بیشتری دارد.

گرانترین محله و سرمایه گذاری پرریسک و پرسود «شیخ زاهد» در لاهیجان با میانگین قیمت بالا نشان می دهد که سرمایه گذاری در این محله با قیمت ورودی بالایی همراه است اما به دلیل احتمال رشد و عرضه کم جزو بهترین ها دسته بندی شده است.

برای سرمایه گذاری با بودجه محدود محله های رشت و امامزاده بندرانزلی مناسب هستند. و این محله ها با وجود قیمت پایین تر به دلیل داشتن شرایط رشد پایدار و عرضه کم در لیست برتر قرار گرفته اند.

تحلیل منطقه ای : بازار ملک در استان گیلان (رشت، لاهیجان، بندرانزلی) به دلیل داشتن سه شهر از چهار شهر پویاترین و سودآورترین منطقه برای سرمایه گذاری ملک به شمار می رود.

تحلیل کد: در این بخش داده ها را آماده سازی و براساس فیلترین گروه بندی و محاسبه شاخص های اصلی انجام شده و امتیاز سرمایه گذاری با استفاده از فرمول زیر حساب شده است. که با توجه به استراتژی (ارزش بالا + عرضه کم) بهترین پتانسیل سرمایه گذاری شناسایی شد.

$$InvestmentScore = (\frac{AveragePrice}{\text{Maximum Average Price}}) \times (\frac{1}{AdCount})$$

تحلیل آمار ساخت و ساز جدید و قیمت بالا:

در این بخش از تحلیل هدف این است که برای سرمایه گذاری روی خانه های گران قیمت سرمایه گذاری کنیم یا ارزان قیمت.

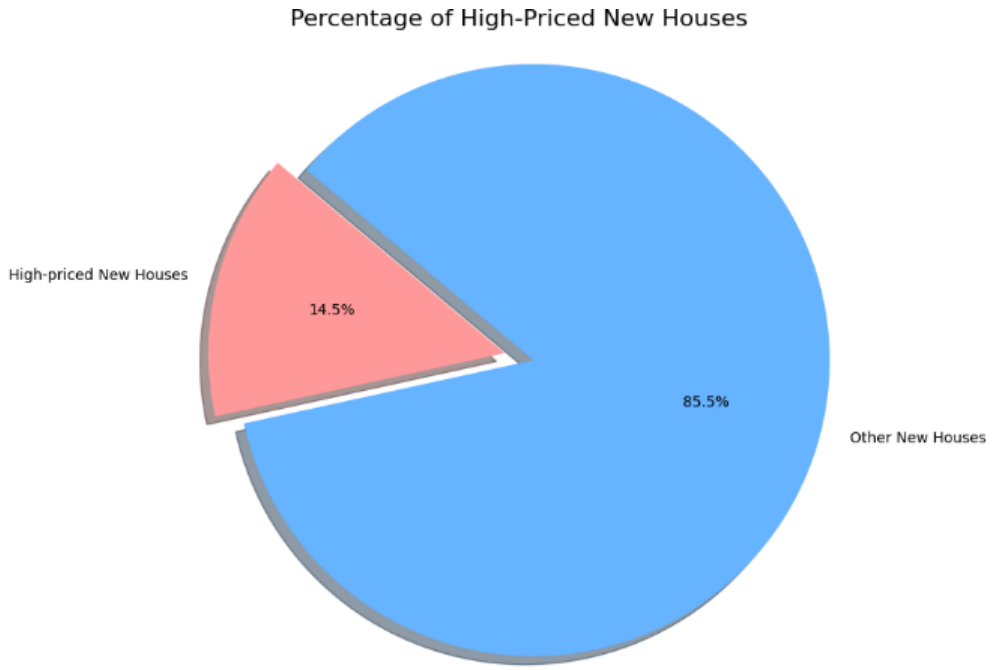
```
Considering recent construction years from 1399 to 1403.
The 75th percentile (Q3) for 'price_per_meter' is: 45.925925925925924
Number of houses built in recent years (1399-1403): 260858
Number of high-priced houses among these: 37922
Percentage of high-priced new houses: 14.54%
```

این آمار نشان می‌دهد که چگونه سهم املاک لوکس و گران‌قیمت در ساخت و سازهای جدید توزیع شده است.

از آنجایی که Q3(چارک سوم) به طور کلی ۲۵ درصد بالای بازار را جدا می‌کند، انتظار می‌رفت که حدود ۲۵٪ از کل داده‌های موجود در این دسته قرار بگیرند. در میان ۲۶۰,۸۵۸ خانه ساخته شده در سال‌های ۱۳۹۹ تا ۱۴۰۳، تنها ۱۴.۵۴٪ (حدود نیمی از ۲۵٪ مورد انتظار) جزو خانه‌های گران‌قیمت بالاتر از Q3هستند.

نتیجه گیری که احتمالا از این بخش خواهیم داشت این است که تمرکز ساخت و ساز در سال های اخیر بیشتر بر روی خانه های با قیمت متوسط متمرکز شده و یا کمبود عرضه لوکس و گران قیمت نسبت به سهم آن در کل بازار کاهش یافته است که نشان از کاهش تمایل سازندگان به سرمایه گذاری در بخش لوکس می باشد.

در نمودار زیر سهم غالب از ساخت و سازهای پنج سال اخیر در دسته قیمت های متوسط و ارزان نشان داده شده است.



ذینفع 3- طراحان داخلی :

تحلیل داده برای ذینفع 2: طراحان داخلی : شناسایی بازارهای هدف سود آور: با استفاده از این روش تحلیلی، می‌توانیم شهرها و مناطقی را شناسایی کنیم که بالاترین پتانسیل برای ارائه خدمات طراحی داخلی را دارند. این تحلیل بر اساس داده‌های واقعی املاک انجام می‌شود و به شما کمک می‌کند تصمیم‌گیری دقیق‌تر و هدفمندتری داشته باشید.

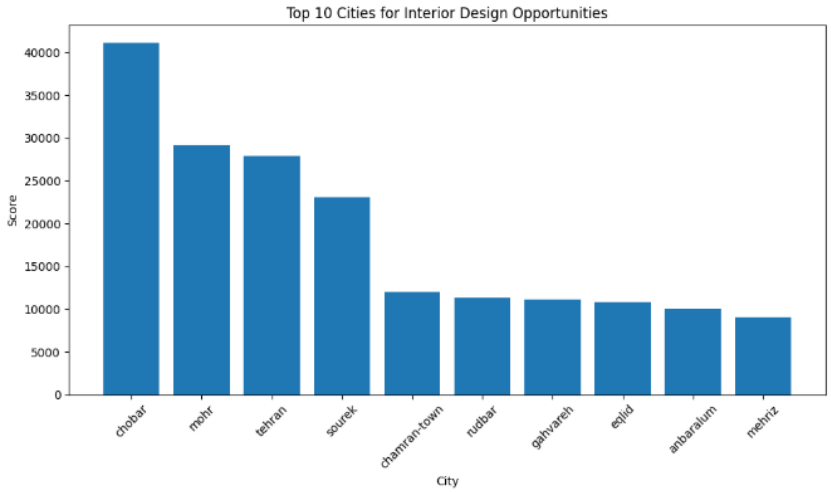
ه کمک این روش می‌توانید:

- تشخیص دهید کدام شهرها **بیشترین ظرفیت و تقاضا** برای پروژه‌های طراحی داخلی را دارند.
- بفهمید در کدام مناطق **تعداد بازسازی و نوسازی بیشتر** است (مشتریان آماده همکاری).
- شناسایی کنید کدام مناطق **بیشترین فایل پیش‌فروش** را دارند (ساختمان‌های تازه‌ساز).
- بررسی کنید در کدام مناطق **میانگین متراژ واحدها بالاتر** است (پروژه‌های بزرگ‌تر و سودآورتر).

تحلیل کد: کد سه نوع داده مهم را در سطح city_slugگروه‌بندی و محاسبه می‌کند (rebuilt_count, pre_sell_count, avg_building_size). که هر کدام به ترتیب نشان دهنده (پروژه های بزرگ تر و سود آورتر، بیشترین پروژه های تازه ساز که نیاز به طراحی دارند، بالاترین تقاضا برای بازسازی و پروژه های مدرن سازی) است. که در این مرحله، یک امتیاز ترکیبی برای هر شهر محاسبه می شود.

$$Score = (RebuiltCount \times 0.6) + (Pre - sellCount \times 0.3) + (AverageBuildingSize \times 0.1)$$

و خروجی این فایل 10 شهر برتر براساس ترکیبی تقاضای بازسازی، ساخت و ساز جدید و اندازه پروژه ها، بیشترین پتانسیل را برای کسب و کار طراحی داخلی مشخص می کند.



با توجه به مطالب بالا، این نمودار 10 شهر برتر که پتانسیل کلی بازار برای خدمات طراحی داخلی را نشان می دهد، رتبه بندی کرده است.

تفسیر: **شهر چوبار (chobar)** با فاصله بسیار زیادی نسبت به سایر شهرها در صدر قرار دارد و این موضوع نشان‌دهنده بزرگترین حجم تقاضا، فعالیت‌های بازسازی و پروژه‌های جدید در این شهر است.

این نمودار نشان می دهد (مهر و چوبار) در رده بالاتری قرار دارند . اگرچه این شهرها شاید به اندازه تهران شناخته شده نباشند، اما امتیاز بالای آن ها نشان دهنده فعالیت بازار ملک بالا و متمرکز است که می تواند ناشی از پروژه های بزرگ توسعه یافته باشد.

شهرهای میانی (شهر سورک، چمران، رودبار، گهواره ..) هستند و حضور شهری مانند **تهران** در رتبه سوم نشان می‌دهد که اگرچه پایتخت حجم آگهی‌های بالایی دارد، اما **امتیاز سودآوری** (بر اساس معیارهای بازسازی و متراژ) در شهرهای دیگر (چوبار و مهر) بالاتر است.

تصمیم گیری برای جذب مشتری و انتخاب محل دفتر با داده های واقعی و عددی شانس بیشتری برای استراتژی هوشمندانه برای افزایش سود و بهره وری کسب و کار مشاور املاک خواهد داد.

ذینفع4- مشاورین املاک:

این تحلیل استراتژیک بر اساس محاسبه امتیاز سودآوری، مستقیماً به شما به عنوان مشاور املاک کمک می‌کند تا بیشترین بازدهی را از فعالیت‌های خود کسب کنید.

با استفاده از این تحلیل، شما قادر خواهید بود:

- شناسایی دفاتر پربازده :مشخص می‌کنید کدام شهرها و محله‌ها بیشترین سود عملیاتی را برای راه‌اندازی و تمرکز فعالیت‌های دفتر مشاور املاک شما دارند.
- تمرکز بر پروژه‌های بزرگ :مناطقی که ترکیبی از قیمت فروش بالا (معمولاً درصد کمیسیون بیشتر)، واحدهای بزرگ (پروژه‌های کلان‌تر) و تعداد معاملات فعال (حجم تراکنش بالا) دارند، به عنوان مناطق بهینه برای تمرکز فعالیت شناسایی می‌شوند.
- انتخاب محل دفتر مبتنی بر داده :تصمیم‌گیری شما برای جذب مشتری، تخصیص منابع و انتخاب محل دفتر با داده‌های واقعی و عددی پشتیبانی می‌شود، نه حدس و گمان.

تحلیل سودآوری این امکان را می دهد که استراتژی هوشمندانه ای برای افزایش سود و بهره وری کسب و کار مشاور املاک خود اجرا کنید.

تحلیل کد: ستون های کلیدی انتخاب می شوند و براساس جفت شهر و محله گروه بندی شده اند، امتیاز سود برای هر محله با استفاده از فرمول زیر که شبیه ساز پتانسیل درآمدی برای مشاور املاک است محاسبه شده است

$$ProfitScore = AvgPricePerMeter \times AvgSize \times DealCount$$

و با استفاده از تابع idxmax فقط محله ها ای را انتخاب می کند که بالاترین profit_score را در هر شهر کسب کرده است.

این نمودار به مشاوران املاک کمک می‌کند تا به جای تمرکز بر شهرهای بزرگ عمومی، فعالیت خود را بر روی محله‌های خاص و سودآور متمرکز کنند:

- **مشهد (بونارستان) :**بالاترین امتیاز کلی را دارد. این منطقه برای کسب **کمیسیون‌های بزرگ** از املاک گران‌قیمت و بزرگ مناسب است.
- **قم (زنبیل‌آباد) :**دارای یک **بازار فعال و پرسود** است که با حجم معاملات بالا (۴۱۷ معامله) جریان نقدی قوی را تضمین می‌کند.
- **کرج (کمال‌شهر) :**با ۷۲۱ معامله، **بالاترین حجم معاملات** در کل داده‌ها را دارد، اما به دلیل قیمت/متراژ پایین، برای استراتژی **حجم فروش بالا** مناسب است.
- **لاهیجان و بندرانزلی :**این شهرها با تعداد معاملات بسیار کم (۷ و ۵ معامله) و امتیاز سودآوری پایین، دارای اعتبار آماری کمتری هستند و پتانسیل درآمدزایی پایین‌تری را نشان می‌دهند.

5- خوشه بندی آگهی های املاک :

پیش پردازش داده : در این مرحله با استفاده از کد پیش پردازش داده را برای آماده سازی یک مجموعه داده ملکی بزرگ جهت استفاده در خوشه بندی (مانند الگوریتم k-means) انجام داده ایم.

هدف نهایی این فرآیند تبدیل داده های خام و متنوع (اعداد، متن و مقادیر بولی) به یک فرمت استاندارد و قابل استفاده توسط الگوریتم های یادگیری ماشین است. این فرآیند شامل تبدیل ویژگی های غیرعددی، مدیریت داده های از دست رفته (NaN) و استانداردسازی مقادیر است. و نتیجه نهایی این بخش آماده تحویل به الگوریتم های خوشه بندی است و تمام ویژگی ها اکنون مقیاس بندی شده اند.

```
Preprocessing complete: Rooms count converted, Boolean columns converted, N
aNs handled, categorical features one-hot encoded, and numerical features s
caled.
Shape of final DataFrame for clustering: (729385, 462)
```

کاهش ابعاد :(Dimensionality Reduction) از آنجایی که داده‌های نهایی دارای صدها ویژگی هستند (طبق خروجی پیش‌پردازش: ۷۲۹,۳۸۵ سطر و ۴۶۲ ستون)، برای ترسیم نتایج، احتمالاً از تکنیک‌هایی مانند **PCA**یا **t-SNE** برای کاهش ابعاد به ۲ یا ۳ بُعد استفاده خواهد شد.

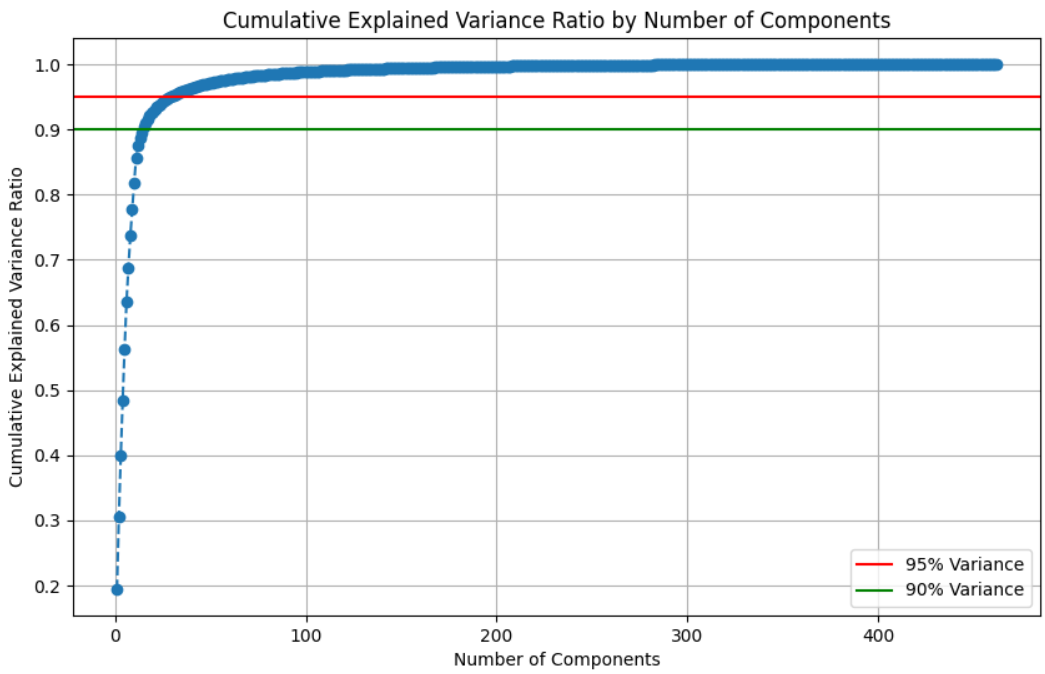
به عنوان مثال ستون هایی مانند has_elevator از مقادیر متنی به 0,1 تبدیل شده اند و مقادیر از دست رفته در ستون های عددی با میانه و در ستون های دسته ای با مد پر شدند. و ویژگی های کلیدی(مانند price_per_meter, building_size) استانداردسازی شدند تا هیچ یک از ویژگی‌ها به دلیل دامنه بزرگتر، بر عملکرد الگوریتم خوشه‌بندی تأثیر نامتناسبی نداشته باشد.

و به این صورت داده ها اکنون آماده اند برای تعیین تعداد بهینه خوشه ها و اجرای الگوریتم خوشه بندی برای شناسایی گروه های املاک در بازار.

کاهش ابعاد با استفاده از PCA

در ابتدا ما اطلاعی از تعداد خوشه ها نداریم و بنابراین از متدهای مختلف استفاده می کنیم تا خوشه بندی بهینه را شناسایی کنیم و ما به دنبال نقطه ای هستیم که نمودار به آرامی شیب خود را از دست می دهد

در این فرآیند حذف نویز که PCA مولفه هایی با واریانس پایین (نویز) را حذف می کنیم و بهبود خوشه بندی که با کاهش ابعاد سرعت پردازش الگوریتم های خوشه بندی را افزایش داده

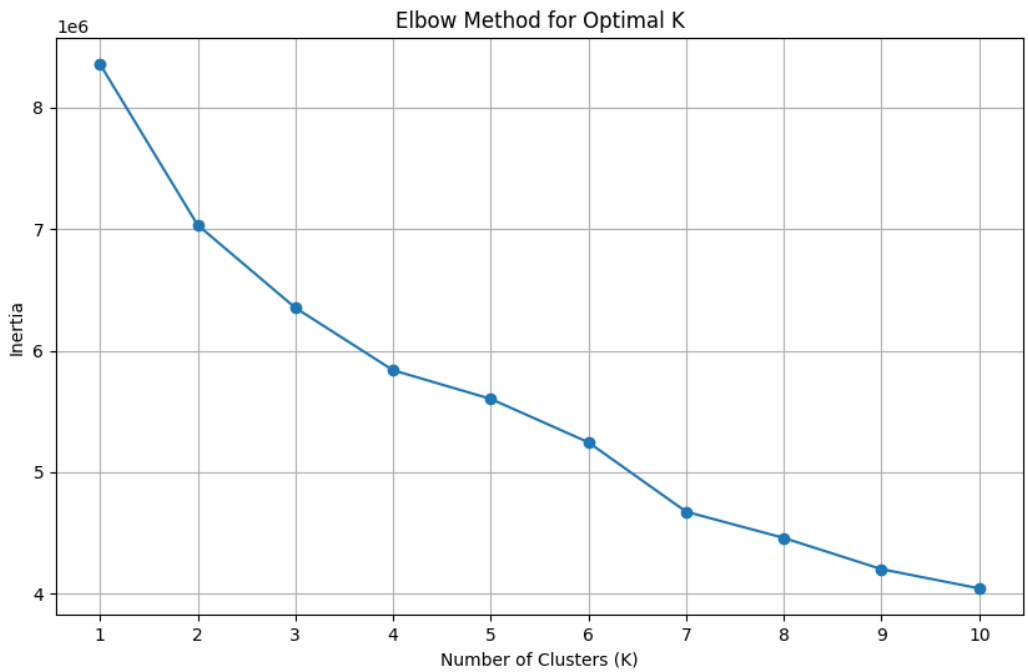



```
Number of components to explain 90% variance: 15
Number of components to explain 95% variance: 30
Selected number of components for PCA: 15
PCA dimensionality reduction complete.
Shape of original data: (729385, 462)
Shape of PCA transformed data: (729385, 15)
```

در این مرحله ما موفق شدیم تعداد ویژگی های مجموعه داده را به 15 مولفه کاهش دهیم. این کاهش به شدت سرعت محاسبات در خوشه بندی را افزایش می دهد و به الگوریتم ها کمک می کند تا الگوهای پنهان را آسان تر بیابند. و با انتخاب 15 مولفه می توانیم 90٪ از واریانس کلی را حفظ کنیم. و این نتایج احتمالا خوشه بندی شفاف تر و سریع تری را ارائه خواهند داد.

خوشه بندی

روش :K-Means الگوریتم K-Means برای گروه‌بندی داده‌ها به K خوشه استفاده می‌شود. کلید موفقیت این الگوریتم، انتخاب درست مقدار K است. در ادامه این کد از معیار اینرسی (Inertia) استفاده می‌کند که مجموع مربعات فاصله بین هر نقطه داده و مرکز خوشه خود است. هدفK-Means ، کمینه‌سازی این مقدار است. محور افقی تعداد خوشه‌ها (K) و محور عمودی میزان اینرسی (Inertia) را نشان می‌دهد. اینرسی، معیاری برای سنجش نزدیکی نقاط داده به مرکز خوشه خود است.



تشخیص K بهینه :با افزایش K، مقدار Inertia همیشه کاهش می‌یابد .نقطه آرنج (Elbow Point) در نمودار، جایی است که شیب کاهش Inertiaبه طور قابل توجهی کند می‌شود. این نقطه به عنوان بهترین مصالحه بین کمیت خوشه‌ها و کیفیت خوشه‌بندی، بهینه‌ترین Kدر نظر گرفته می‌شود.

کاهش شیب اولیه : با توجه به نمودار با افزایش K از ۱ به ۲ و سپس به ۳، مقدار اینرسی کاهش شدیدی را تجربه می‌کند (از حدود ۸.۴ میلیون به حدود ۶.۳ میلیون). این نشان می‌دهد که تقسیم‌بندی املاک به ۲ یا ۳ گروه، بهبود قابل توجهی در کیفیت خوشه‌بندی ایجاد می‌کند. همچنین K=4 به عنوان (Elbow Point)در نظر گرفته می‌شود. پس از K=4، شیب کاهش اینرسی کندتر شده و کاهش ارزش Inertia در ازای افزودن خوشه‌های بیشتر، توجیه کمتری پیدا می‌کند (کاهش سوددهی مشاهده می‌شود). و انتخاب K=4 به این معنی است که بازار مسکن می‌تواند به طور موثر و معنی‌داری به چهار دسته (خوشه) اصلی از املاک تقسیم شود که هر کدام ویژگی‌های متمایز خود (مانند قیمت، متراژ، شهر و امکانات) را دارند. این چهار دسته، نماینده چهار بخش استراتژیک از بازار هستند (برای مثال، لوکس، میان‌رده، ارزان و سرمایه‌گذاری).

گام اول تعیین تعداد بهینه خوشه : کد Elbow Method را برای تعیین بهترین تعداد خوشه ها (K) در داده های کاهش یافته PCA اجرا می کنیم. براساس این روش تعداد بهینه خوشه ها می توانند K=4 یا K=5 باشد که برای تایید این انتخاب و سنجش تراکم خوشه ها اجرای مرحله بعدی یعنی محاسبه امتیاز **Silhouette ضروری است. در این روش، بزرگترین امتیاز سیلوئت، نشان‌دهنده بهینه‌ترین مقدار K است.**

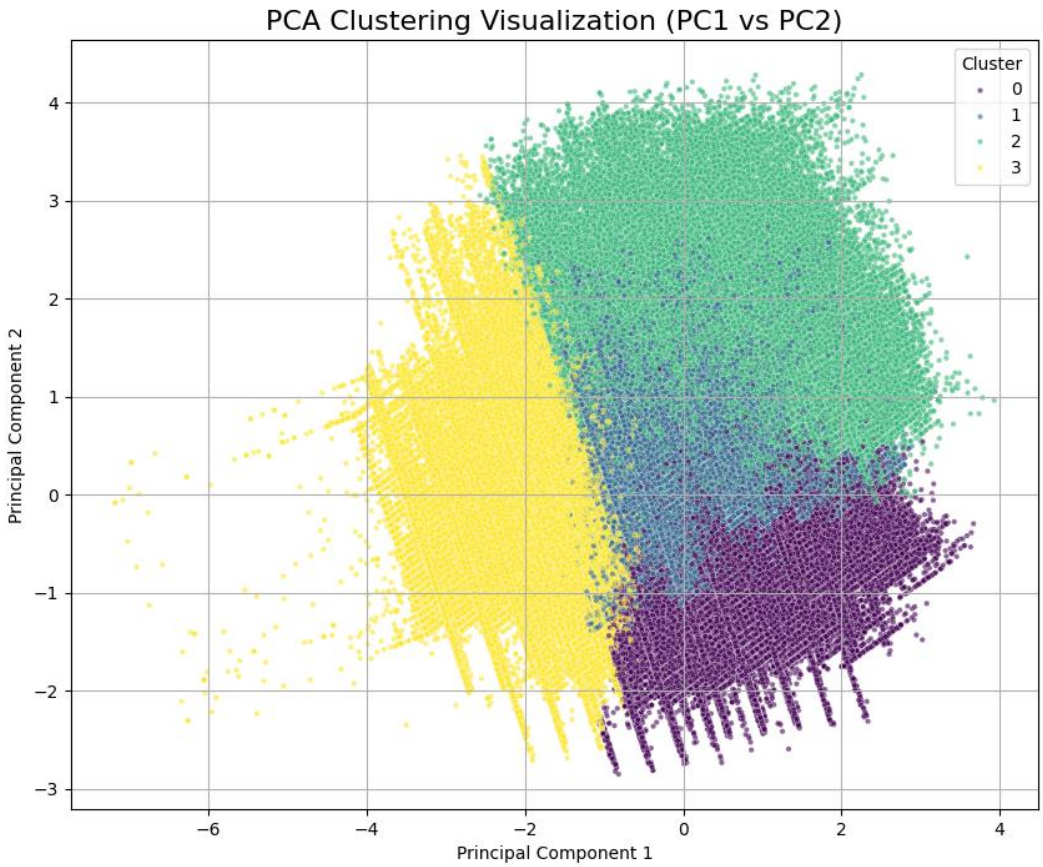
اجرای خوشه بندی K-Means با K=4

توضیح عملیات انجام شده

1. اجرای مدل : ما K=4 را برای اجرای نهایی الگوریتم k-means انتخاب می کنیم. random_state=42برای تضمین قابلیت تکرارپذیری نتایج استفاده می‌شود.

2. خروجی: مدل یک آرایه از برچسب‌ها (مثلاً **0و1و2و3** تولید می‌کند که نشان می‌دهد هر ملک به کدام خوشه تعلق دارد.

3. تفسیرپذیری:



تفسیر خوشه های تحلیل شده PCA: چهار بخش متمایز بازار املاک

بر اساس تحلیل ویژگی‌های عددی و دسته‌ای در هر خوشه، می‌توان بخش‌های مختلف بازار املاک را به شکل زیر تعریف کرد:

- **خوشه 0: آپارتمان‌های میان‌رده شهری (با تمرکز بر مشاوران املاک)**

- **ویژگی‌های عددی :**قیمت متوسط هر متر، اجارهٔ کم هر متر (احتمالاً به دلیل تمرکز بیشتر بر املاک فروش یا بازده اجاره کم نسبت به قیمت)، متراژ نسبتاً کوچک (حدود ۴۰۰ متر)، ۱–۲ اتاق، سال ساخت قدیمی‌تر (حدود ۱۳۹۰). حضور زیاد امکانات پایه مثل بالکن، آسانسور، انباری و پارکینگ، اما امکانات لوکس بسیار کم.
- **ویژگی‌های دسته‌ای :**عمدتاً در تهران، کرج، مشهد. تمرکز بر «مسکونی-فروش» و «مسکونی-اجاره». غالباً سند «تک‌برگی». عمدتاً توسط «مشاور املاک» ثبت می‌شوند. جهت‌های شمال/جنوب و کف‌پوش سرامیک رایج است. نوع ملک بیشتر «سایر» و سپس «جنگلی.«
- **تفسیر :**این خوشه نمایانگر املاک مسکونی میان‌رده در شهرهای بزرگ است که عمدتاً آپارتمانی و توسط مشاوران املاک مدیریت می‌شوند. اختلاف کم اجاره نسبت به قیمت فروش ممکن است نشان‌دهنده‌ی بازار خرید فعال‌تر یا نبود داده‌های قوی در بخش اجاره باشد.

- **خوشه 1: ترکیبی تجاری/مسکونی (نوسان زیاد قیمت و اجاره)**

- **ویژگی‌های عددی :**قیمت متوسط هر متر، اما/*اجارهٔ بسیار بالا* (۱۳۳۸۰۴ که احتمال وجود نقاط پرت بسیار بالا یا املاک خاص تجاری/کوتاه‌مدت را نشان می‌دهد). بزرگ‌ترین متراژها (حدود ۹۰۰۰ متر)، اتاق کم (حدود ۱)، سال ساخت جدیدتر (۱۳۹۲). کم‌ترین حضور امکانات پایه.
- **ویژگی‌های دسته‌ای :**پراکندگی در شهرهای بزرگ (تهران، مشهد، اصفهان). ترکیبی از فروش/اجاره مسکونی و تجاری و همچنین اجاره کوتاه‌مدت. حضور زیاد «زمین-قدیمی» و «مغازه-اجاره». تنوع در نوع سند. غالباً توسط مشاور املاک ثبت می‌شود. کف‌پوش بیشتر سرامیک، سپس موکت و موزاییک.
- **تفسیر :**این خوشه ترکیبی از املاک بزرگ، زمین‌های بایر، و املاک تجاری با نرخ اجارهٔ بسیار متغیر است. وجود اجاره‌های فوق‌العاده بالا احتمالاً ناشی از قراردادهای تجاری خاص یا داده‌های پرت می‌باشد.

- **خوشه 2: آپارتمان‌های لوکس شهری (با تمرکز بر مشاوران املاک)**

- **ویژگی‌های عددی :**قیمت متوسط هر متر (مشابه خوشه 0) اما تعداد اتاق‌ها بیشتر (۲–۳)، سال ساخت جدیدتر (۱۳۹۶)، حضور زیاد امکانات پایه. اجاره هر متر متوسط اما بیشتر از خوشه 0 و بسیار کمتر از خوشه 1.
- **ویژگی‌های دسته‌ای :**مشابه خوشه 0 اما تمرکز بیشتر بر املاک مسکونی فروش/اجاره در آپارتمان‌ها و ویلاها. سند تک‌برگی، ثبت توسط مشاور املاک، جهت‌های شمال/جنوب، کف‌پوش عمدتاً سرامیک و سپس سنگ. نوع ملک «جنگلی» بسیار زیاد) ٪۶۷.(
- **تفسیر :**این خوشه نمایانگر املاک مسکونی لوکس‌تر و نسبتاً جدید در شهرهای بزرگ است. تعداد اتاق بیشتر، سال ساخت جدید و امکانات بالا، نشان‌دهنده‌ی جذابیت بیشتر برای خریداران و مستأجران است.

- **خوشه 3: بازار ویلا/زمین شمال (پر از امکانات رفاهی)**

- **ویژگی‌های عددی :**کم‌ترین قیمت هر متر، کم‌ترین اجاره هر متر، متراژ زیاد (حدود ۳۱۴۰ متر)، بیشترین تعداد اتاق (۲–۳)، جدیدترین سال ساخت (۱۳۹۹). حضور بالا در امکانات لوکس مثل نگهبانی، باربیکیو، استخر، جکوزی و سونا. تقریباً ٪۱۰۰ دارای آب، برق، گاز.

- **ویژگی‌های دسته‌ای :**متمرکز در شهرهای شمالی: چمستان، نور، محمودآباد، نوشهر، رامسر. تمرکز زیاد بر «مسکونی-فروش» و «اجاره موقت» ویلا. سند تک‌برگی، ثبت توسط مشاور املاک. کف‌پوش سرامیک. نوع ملک «جنگلی» و سپس «ساحلی.«

- **تفسیر :**این خوشه دقیقاً نشان‌دهنده بازار ویلاهای شمال و املاک گردشگری است. امکانات لوکس، متراژ بالا و سال ساخت جدید، این املاک را به گزینه‌هایی مناسب اقامت تفریحی یا سرمایه‌گذاری تبدیل می‌کند.

مقایسه با انتظارات:

نمایش تصویری خوشه‌ها نشان می‌دهد که تفکیک مناسبی به‌خصوص در خوشه 3 وجود دارد. روش‌های Elbow و Silhouette مقدار ۳ یا ۴ را پیشنهاد می‌کردند و K=4 توانسته تقسیم‌بندی معناداری ایجاد کند. تفسیرها نیز با انتظار ما از بازار املاک، از نظر موقعیت، نوع ملک، سن بنا، امکانات و قیمت سازگار است.

خلاصه:

یافته‌های کلیدی تحلیل داده

- **کارآمدی کاهش ابعاد (PCA)**: استفاده از تحلیل مؤلفه‌های اصلی به‌طور مؤثر ابعاد داده را کاهش داد و امکان نمایش واضح تفکیک خوشه‌ها در یک فضای دوبعدی را فراهم کرد. نمودار پراکندگی PC1 در برابر PC2 گروه‌بندی‌های مشخصی را نشان داد که تأیید می‌کند مقدار K=4 تفکیک‌های معناداری در بازار ایجاد کرده است.

- **خوشه 0: آپارتمان‌های میان‌رده شهری:** این بخش شامل املاک مسکونی، عمدتاً آپارتمان‌ها، در شهرهای بزرگی مانند تهران و کرج است. ویژگی‌های آن شامل قیمت متوسط، بازده اجاره پایین‌تر نسبت به قیمت فروش، متراژ حدود ۴۰۰ متر، ۱ تا ۲ اتاق و سال ساخت قدیمی‌تر (حدود ۲۰۱۱) است. این املاک معمولاً دارای امکانات پایه بوده و عمدتاً توسط مشاوران املاک مدیریت می‌شوند.

- **خوشه 1: املاک تجاری/مسکونی متنوع با نوسان شدید اجاره:** این خوشه نشان‌دهنده مجموعه‌ای از املاک بزرگ، احتمالاً شامل فضاهای تجاری یا زمین‌های بایر است، با میانگین اجاره هر متر بسیار بالا (۱۳۳٬۸۰۴ دلار). این املاک بیشترین متراژ متوسط (حدود ۹۰۰۰ متر)، تعداد اتاق کم (حدود ۱) و سال ساخت جدیدتر (حدود ۲۰۱۳) دارند. این تنوع نشان‌دهنده بازاری با تمرکز سرمایه‌گذاری بالا و فعالیت‌های تخصصی تجاری است.

- **خوشه 2: آپارتمان‌های لوکس شهری:** این بخش شامل آپارتمان‌ها و خانه‌های مسکونی جدیدتر و سطح‌بالا در شهرهای بزرگ مانند تهران، مشهد و کرج است. آن‌ها دارای تعداد اتاق بیشتر (۲–۳)، سال ساخت جدیدتر (حدود ۲۰۱۷)، و مجموعه کامل امکانات پایه هستند. این ویژگی‌ها نشان‌دهنده جذابیت بیشتری نسبت به خوشه 0 بوده و غالباً توسط مشاوران املاک مدیریت می‌شوند.

- **خوشه 3: بازار ویلا/زمین شمال:** این خوشه عمدتاً در شهرهای شمالی و گردشگری قرار دارد و شامل ویلاها، خانه‌های تفریحی یا زمین است. این املاک دارای جدیدترین سال ساخت (حدود ۲۰۲۰)، بیشترین تعداد اتاق (۲–۳)، و متراژ بالاتر (حدود ۳۱۴۰ متر) هستند. وجود امکانات لوکس مانند استخر، جکوزی، سونا و نگهبانی از ویژگی‌های بارز این خوشه است، که معمولاً در دسته‌های فروش مسکونی یا اجاره موقت ارائه می‌شوند.

بینش‌ها و گام‌های بعدی

- بخش‌های مختلف شناسایی‌شده، پایه محکمی برای استراتژی‌های بازاریابی هدفمند و برنامه‌ریزی توسعه فراهم می‌کنند و به ذینفعان امکان می‌دهند روی نیازهای مشتریان و انواع خاصی از املاک تمرکز کنند (مثلاً ویلاهای لوکس شمال در مقابل آپارتمان‌های میان‌رده شهری).

- نیاز به بررسی بیشتر درباره اجاره‌های بسیار بالا در خوشه 1 وجود دارد تا مشخص شود این نرخ‌ها ناشی از قراردادهای تجاری تخصصی، اجاره‌های کوتاه‌مدت یا داده‌های پرت هستند؛ زیرا این موارد می‌توانند تأثیر چشمگیری بر ارزش‌گذاری بازار و تصمیمات سرمایه‌گذاری داشته باشند.

مقایسه با انتظارات:

نمایش تصویری خوشه‌ها نشان می‌دهد که تفکیک مناسبی به‌خصوص در خوشه 3 وجود دارد. روش‌های Elbow و Silhouette مقدار ۳ یا ۴ را پیشنهاد می‌کردند و K=4 توانسته تقسیم‌بندی معناداری ایجاد کند. تفسیرها نیز با انتظار ما از بازار املاک، از نظر موقعیت، نوع ملک، سن بنا، امکانات و قیمت سازگار است

خلاصه:

یافته‌های کلیدی تحلیل داده

- **کارآمدی کاهش ابعاد :(PCA)** استفاده از تحلیل مؤلفه‌های اصلی به‌طور مؤثر ابعاد داده را کاهش داد و امکان نمایش واضح تفکیک خوشه‌ها در یک فضای دوبعدی را فراهم کرد. نمودار پراکندگی PC1 در برابر PC2 گروه‌بندی‌های مشخصی را نشان داد که تأیید می‌کند مقدار K=4 تفکیک‌های معناداری در بازار ایجاد کرده است.
- **خوشه 0: آپارتمان‌های میان‌رده شهری :**این بخش شامل املاک مسکونی، عمدتاً آپارتمان‌ها، در شهرهای بزرگی مانند تهران و کرج است. ویژگی‌های آن شامل قیمت متوسط، بازده اجاره پایین‌تر نسبت به قیمت فروش، متراژ حدود ۴۰۰ متر، ۱ تا ۲ اتاق و سال ساخت قدیمی‌تر (حدود ۲۰۱۱) است. این املاک معمولاً دارای امکانات پایه بوده و عمدتاً توسط مشاوران املاک مدیریت می‌شوند.
- **خوشه 1: املاک تجاری/مسکونی متنوع با نوسان شدید اجاره :**این خوشه نشان‌دهنده مجموعه‌ای از املاک بزرگ، احتمالاً شامل فضاهای تجاری یا زمین‌های بایر است، با میانگین اجاره هر متر بسیار بالا (۱۳۳،۸۰۴ دلار). این املاک بیشترین متراژ متوسط (حدود ۹۰۰۰ متر)، تعداد اتاق کم (حدود ۱) و سال ساخت جدیدتر (حدود ۲۰۱۳) دارند. این تنوع نشان‌دهنده بازاری با تمرکز سرمایه‌گذاری بالا و فعالیت‌های تخصصی تجاری است.
- **خوشه 2: آپارتمان‌های لوکس شهری :**این بخش شامل آپارتمان‌ها و خانه‌های مسکونی جدیدتر و سطح‌بالا در شهرهای بزرگ مانند تهران، مشهد و کرج است. آن‌ها دارای تعداد اتاق بیشتر (۲–۳)، سال ساخت جدیدتر (حدود ۲۰۱۷)، و مجموعه کامل امکانات پایه هستند. این ویژگی‌ها نشان‌دهنده جذابیت بیشتری نسبت به خوشه 0 بوده و غالباً توسط مشاوران املاک مدیریت می‌شوند.
- **خوشه 3: بازار ویلا/زمین شمال :**این خوشه عمدتاً در شهرهای شمالی و گردشگری قرار دارد و شامل ویلاها، خانه‌های تفریحی یا زمین است. این املاک دارای جدیدترین سال ساخت (حدود ۲۰۲۰)، بیشترین تعداد اتاق (۲–۳)، و متراژ بالاتر (حدود ۳۱۴۰ متر) هستند. وجود امکانات لوکس مانند استخر، جکوزی، سونا و نگهبانی از ویژگی‌های بارز این خوشه است، که معمولاً در دسته‌های فروش مسکونی یا اجاره موقت ارائه می‌شوند.

بینش‌ها و گام‌های بعدی

- بخش‌های مختلف شناسایی‌شده، پایه محکمی برای استراتژی‌های بازاریابی هدفمند و برنامه‌ریزی توسعه فراهم می‌کنند و به ذینفعان امکان می‌دهند روی نیازهای مشتریان و انواع خاصی از املاک تمرکز کنند (مثلاً ویلاهای لوکس شمال در مقابل آپارتمان‌های میان‌رده شهری).

- نیاز به بررسی بیشتر درباره اجاره‌های بسیار بالا در خوشه 1 وجود دارد تا مشخص شود این نرخ‌ها ناشی از قراردادهای تجاری تخصصی، اجاره‌های کوتاه‌مدت یا داده‌های پرت هستند؛ زیرا این موارد می‌توانند تأثیر چشمگیری بر ارزش‌گذاری بازار و تصمیمات سرمایه‌گذاری داشته باشند.

6- مدل سازی قیمت:

- پیش‌پردازش: در این بخش فاز آماده سازی داده ها برای ساخت مدل پیش بینی قیمت است (متغیر هدف : value_rent_or_price). هدف تبدیل ویژگی های خام به فرمت کاملا عددی برای استفاده در الگویابی یادگیری ماشین (رگرسیون، شبکه های عصبی و ...) است. که بالاتر توضیح داده شده است.

  انتخاب و آماده سازی متغیرهای (X,Y)

  متغیر هدف :(Y) ستون **value_rent_or_price**(شامل قیمت فروش یا ارزش اجاره) به عنوان متغیر وابسته برای پیش‌بینی انتخاب شده است.

  متغیرهای ورودی :(X) متغیرهای ورودی شامل تمامی ستون‌های **عددی اصلی (numeric_cols)** و ستون‌های **دسته‌بندی (categorical_cols)**هستند.

- استانداردسازی و تقسیم داده‌ها

در این بخش مراحل نهایی آماده سازی و استاندارسازی داده های ورودی برای مدل های یادگیری ماشین انجام شده است و سپس مجموعه داده ها را به زیرمجموعه های **آموزش(Train) ، اعتبارسنجی (Validation) و آزمون (Test)** تقسیم می کنیم.

عملیات :با استفاده از StandardScaler، داده‌های ورودی آماده شده برای مدل‌های خطی (X_Enc) **استانداردسازی** می‌شوند.

آماده‌سازی داده برای مدل‌های درختی

دیتافریم ورودی جدیدی به نام X_Treeایجاد می‌شود. هدف این دیتافریم برای مدل‌هایی مانند **درخت تصمیم (Decision Tree)** و **جنگل تصادفی (Random Forest)** آماده شده است. این مدل‌ها به استانداردسازی داده‌ها یا رمزگذاری One-Hot برای ویژگی‌های دسته‌بندی نیاز ندارند، بنابراین از ستون‌های دسته‌بندی غیررمزگذاری شده استفاده می‌کند.

بررسی ابعاد داده

خروجی تأیید می‌کند که هر دو مجموعه ورودی X_Enc) و (X_Treeو مجموعه هدف (y) دارای **۱۲۵,۹۶۰ ردیف** و ۴۴ ویژگی ورودی هستند.

مدلسازی

ا با استفاده از متغیرهای موثر، شش مدل یادگیری ماشین با رویکردهای گوناگون (خطی، همسایگی و یادگیری گروهی) را برای پیش‌بینی قیمت آگهی‌ها اجرا کرده‌ایم. داده‌های ورودی برای این مدل‌ها **استانداردسازی** و **رمزگذاری (Label Encoded)** شده‌اند.

مدل	نوع رویکرد	توصیف کوتاه
K نزدیکترین همسایگی(KNN)	الگوریتم همسایگی	پیش‌بینی بر اساس میانگین مقادیر ۵ همسایه‌ی نزدیکتر در فضای ویژگی‌ها.
رگرسیون خطی(Linear Regression)	مدل خطی پایه	یافتن رابطه خطی بین متغیرهای ورودی و متغیر هدف.
رگرسیون لاسو(Lasso)	رگرسیون منظم شده(Regularization)	رگرسیون خطی با جریمه L1 برای کاهش اهمیت یا صفر کردن وزن متغیرهای کم‌اثر.
رگرسیون ریج(Ridge)	رگرسیون منظم شده(Regularization)	رگرسیون خطی با جریمه L2 برای کاهش اندازه ضرایب (کاهش بیش‌برازش).
تقویت گرادیان(Gradient Boosting)	یادگیری گروهی(Ensemble)	ساخت یک مدل قوی از ترکیب متوالی درختان تصمیم ضعیف.
XGBoost	یادگیری گروهی(Ensemble)	یک فریم‌ورک کارآمد و بهینه‌سازی شده برای Gradient Boostingفقط تعریف شده و اجرا نشده است

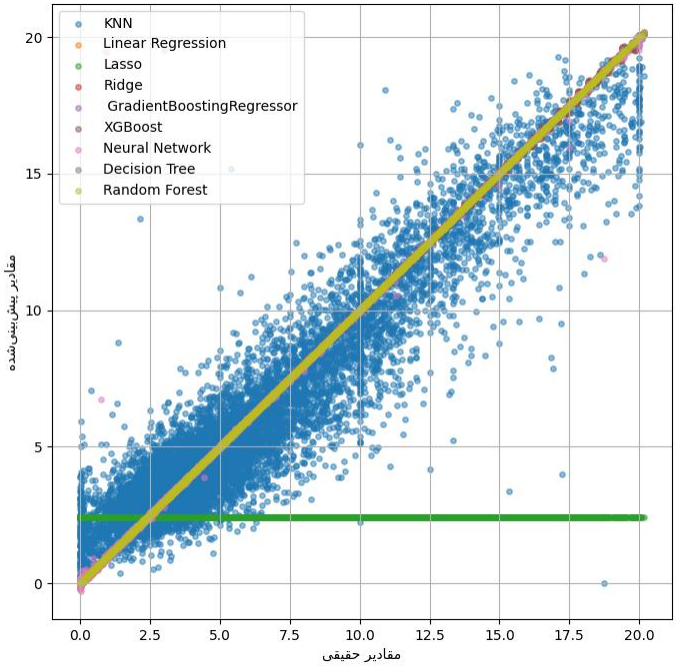
انتخاب بهترین مدل:

مدل‌های **یادگیری گروهی مانند (Gradient Boosting)** و **مدل‌های منظم شده (Ridge/Lasso)** اغلب در مجموعه داده‌های واقعی بهتر از رگرسیون خطی پایه عمل می‌کنند.

در حالت کلی، انتظار می‌رود مدل‌هایی با **پیچیدگی بیشتر** و قابلیت ثبت روابط غیرخطی (مانند **(Gradient Boosting**یا مدل‌هایی که از روابط محلی استفاده می‌کنند **(KNN) خطای کمتری** RMSE/MAE)پایین‌تر سبت به رگرسیون خطی داشته باشند.

نمودار مقایسه مدل‌ها

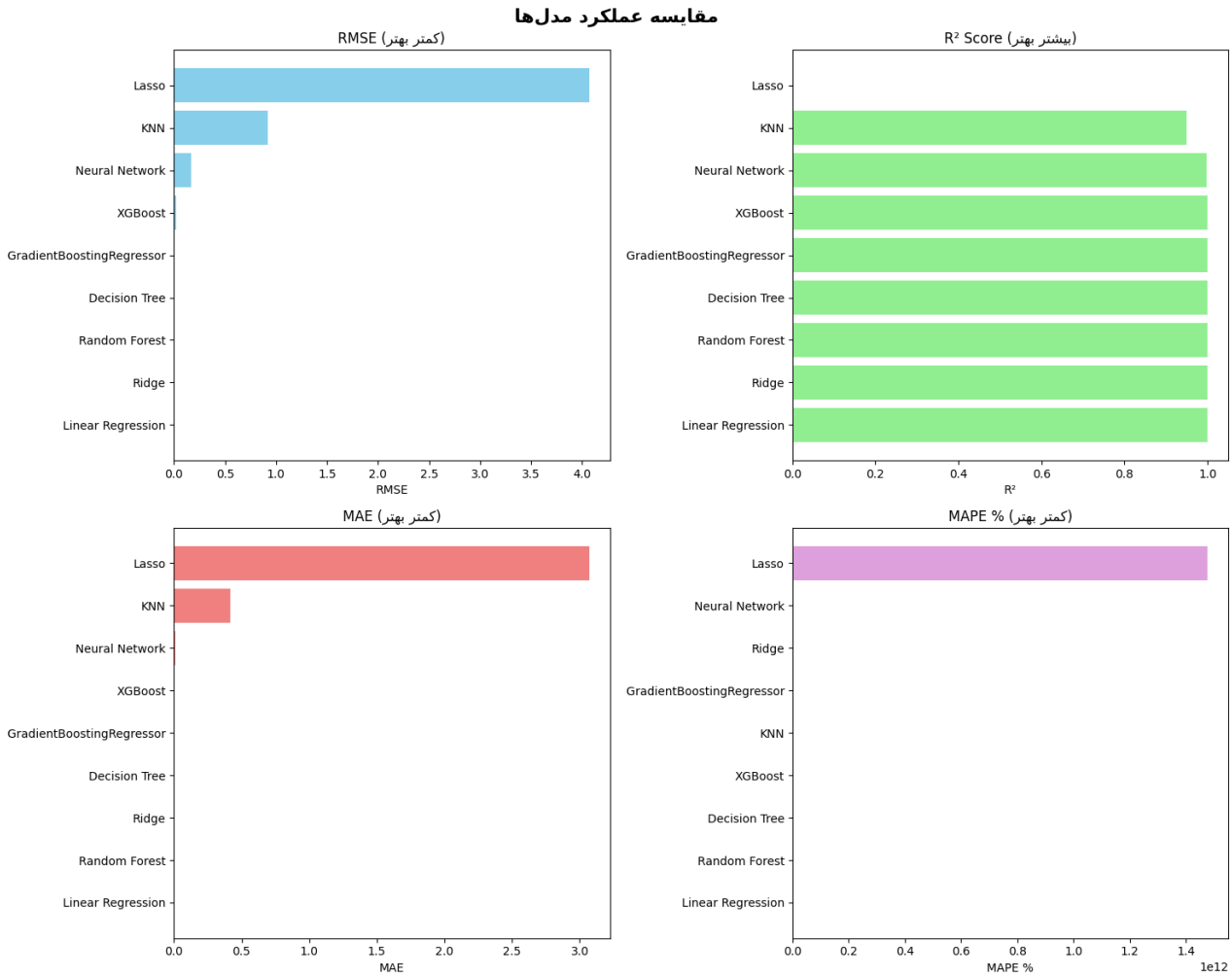
در زیر نمودار پراکندگی (Scatter Plot) برای مقایسه نتایج تمام مدل‌های اجرا شده در برابر مقادیر حقیقی (Actual Values) رسم شده است.



مدل	رنگ	مشاهده و تحلیل عملکرد
KNN، Gradient Boosting، XGBoost، Decision Tree، Random Forest	رنگ‌های فیروزه‌ای، نارنجی، صورتی، زرد (همپوشانی بالا)	**عملکرد عالی :**نقاط این مدل‌ها در یک **نوار باریک** و نزدیک به **خط قطری** (خطی که از مبدأ تا گوشه بالا راست می‌رود) قرار گرفته‌اند. این نشان‌دهنده **دقت بسیار بالای این مدل‌ها** در پیش‌بینی قیمت آگهی‌ها است.
Linear Regression, Lasso, Ridge	سبز (مشترک)	عملکرد ضعیف/نادرست :نقاط این مدل‌ها به شدت از خط قطری منحرف شده‌اند و اغلب در نزدیکی خط افقی $y \sim 2.5$ متمرکز شده‌اند. این یعنی مدل‌های خطی در پیش‌بینی قیمت‌های بالا شکست خورده‌اند و تمایل دارند همه قیمت‌ها را نزدیک به یک مقدار ثابت (میانگین یا میانه پایین) پیش‌بینی کنند.
Neural Network	در پشت سایر نقاط پنهان شده	با توجه به عملکرد ضعیف مدل‌های خطی، اگر مدل شبکه عصبی به درستی تنظیم نشده باشد، عملکرد آن می‌تواند شبیه مدل‌های خطی باشد یا اگر خوب تنظیم شده باشد، با مدل‌های درختی همسو شود.

به عبارت دیگر در انتخاب مدل ، مدل‌های **رگرسیون خطی، لاسو و ریج** به دلیل عدم توانایی در ثبت روابط پیچیده و غیرخطی در داده‌های قیمت ملک، **به طور کامل نامناسب** هستند.

تفسیر نمودار مقایسه عملکرد مدل ها



این چهار نمودار معیار کلیدی ارزیابی برای شش مدل رگرسیون اجرا شده شامل Linear Regression، Ridge، Lasso، KNN، Gradient Boosting وXGBoost را مقایسه می کنند.

با توجه به نمودارها و تفسیر ها **XGBoost**و **GradientBoostingRegressor**به عنوان بهترین مدل‌ها برای پیش‌بینی قیمت آگهی‌ها انتخاب می‌شوند، زیرا موفق شده‌اند تقریباً تمام واریانس داده‌ها (R^2 ~ 1.0) را توضیح دهند و کمترین خطای مطلق و مربعی را ایجاد کنند.

معیار	توضیح	مشاهده در نمودار
RMSE: ریشه میانگین مربع خطا	میزان خطای مدل با جریمه کردن خطاهای بزرگ.	اختلاف بسیار کم در مدل‌های برتر؛ خطای بالای Lasso.
R^2 Score : ضریب تعیین	میزان توضیح واریانس متغیر هدف توسط مدل.	مدل‌های برتر نزدیک به ۱۰۰٪ واریانس را توضیح می‌دهند.
MAE: میانگین خطای مطلق	میانگین فاصله مطلق بین پیش‌بینی و مقدار واقعی.	اختلاف بسیار کم در مدل‌های برتر؛ خطای بالای Lasso.
MAPE %: میانگین خطای مطلق درصد	میانگین خطای پیش‌بینی به صورت درصدی.	مدل‌های برتر خطای بسیار کمی دارند؛ Lasso. خطای بسیار بالای

Choose Best Model

Metric	Goal	Notes
MAE (Mean Absolute Error)	Lower is better	Average absolute error, same unit as target (e.g., price)
RMSE (Root Mean Squared Error)	Lower is better	Penalizes large errors more than MAE
MAPE (Mean Absolute Percentage Error)	Lower is better	Percentage error, good for relative comparison
R^2 (Coefficient of Determination)	Higher is better	Explains how much variance is captured by the model

تحلیل نتایج مدل: دسته بندی قیمت آگهی ها

خروجی تحلیل نشان می‌دهد که مجموعه داده آزمون (Test) به شرح زیر در این سه دسته توزیع شده است:

```
Category
Under-value    19405
Normal          5228
over-value       559
Name: count, dtype: int64
```

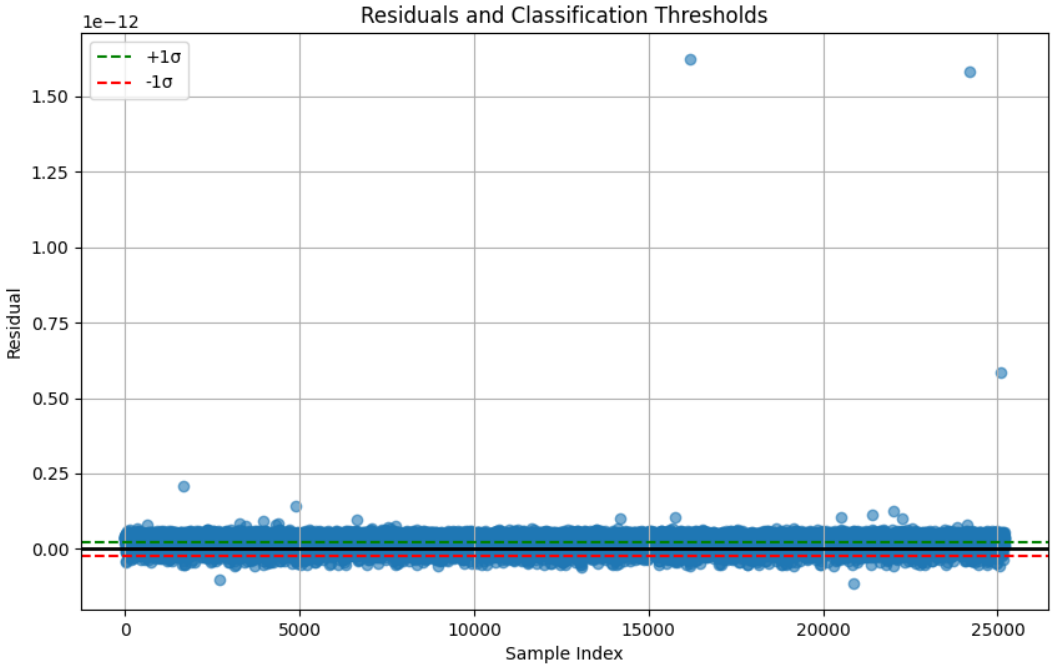
بزرگترین سهم از آگهی‌ها (حدود ۷۷٪ در دسته **Under-value**قرار دارند. این بدان معناست که برای بخش بزرگی از آگهی‌ها، **قیمت واقعی از قیمت پیش‌بینی شده توسط مدل بیشتر است.**

کمترین تعداد آگهی‌ها حدود ۲.۲٪ (در دسته **Over-value** قرار دارند.

مدل با دقت بالا :درصد آگهی‌های "نرمال" (حدود ۲۱٪) نشان می‌دهد که مدل در پیش‌بینی قیمت برای بخش قابل توجهی از بازار، عملکرد دقیق و در محدوده خطای مجاز داشته است. با این حال، غلبه آگهی‌های Under-value می‌تواند نشان‌دهنده یک **سوگیری (Bias)** جزئی در مدل است

تحلیل Plot Residual and Thresholds: شناسایی فرصت های معاملاتی (Under/Over-value)

این نمودار **صحت دسته‌بندی‌های قیمت** (Under-value, Normal, Over-value)را تأیید می‌کند و نشان می‌دهد که مدل ما برای شناسایی آگهی‌های با ارزش معاملاتی بالا (Under-value) بسیار موثر است.



در این بخش آنچه انجام داده ایم قیمت پیش بینی شده مدل را با قیمت واقعی آگهی مقایسه کردیم تا ببینیم یک ملک ارزان است یا گران و از انحراف معیار خطای مدل به عنوان مرز طبیعی خطا استفاده کرده ایم

7- دسته بندی با استفاده از داده های متنی

در این بخش از تحلیل روی استفاده از داده های متنی آگهی (ترکیب عنوان و توضیحات) برای دو هدف زیر در classification تمرکز کردیم

1- پیش بینی دسته بندی ملک (مانند آپارتمان، ویلا، زمین و ..)

2- پیش بینی نوع کاربر آگهی دهنده (مشاور املاک یا شخصی)

در این بخش ستون های title و description با هم ترکیب شده و از تابع cleaning_text برای حذف اعداد، علائم نگارشی و اصلاح فاصله ها استفاده شد. Vectorization برای تبدیل متن به فرمت عددی قابل فهم برای مدل دو روش (BOW و TF-IDF) اعمال شده است.

Bag-of-Words: شمارش دفعات تکرار کلمات

و Term Frequency-Inverse Document Frequency: وزن دهی به کلمات براساس اهمیت آن ها در هر آگهی و کل مجموعه

نتایج ارزیابی مدل های دسته بندی ملک به صورت زیر قابل مشاهده است که مشاهده می شود رگرسیون لجستیک با استفاده از بردارسازی TF-IDF با دقت بیش از 81 درصد بهترین عملکرد را در پیش بینی نوع دسته بندی ملک براساس متن آگهی داشته است. این نشان می دهد که وزن دهی برای این مدل بهتر از صرفا شمارش کلمات BOW عمل کرده است.

Model Evaluation Results:

	Accuracy	Precision	Recall	F1-Score
Multinomial Naive Bayes (BOW)	0.757621	0.773761	0.757621	0.759885
Logistic Regression (TF-IDF)	0.819983	0.815196	0.819983	0.813268
RandomForestClassifier (TF-IDF)	0.786003	0.783649	0.786003	0.775110