به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و

کامپیوتر

درس داده کاوی پیشرفته

تمرین اول

نام و نام خانوادگی	عرفان شهابی
شماره دانشجویی	۸۱۰۱۰۳۱۶۶
تاریخ ارسال گزارش	۱۴۰۳.۱۲.۲۰

فهرست

نمودارها

جدول‌ها

سوال ۱

2. نوع داده ها

داده	پیوسته / گسسته	نوع داده
سال ساخت	گسسته	عددی
مساحت	پیوسته	عددی
طبقه	گسسته	عددی
تعداد کل طبقات	گسسته	عددی
نوع خانه	گسسته	باینری
تعداد اتاق	گسسته	عددی
آسانسور دارد	گسسته	باینری
پارکینگ دارد	گسسته	باینری
سطح امنیت منطقه	گسسته	ترتیبی
نوع پوشش کف واحد	گسسته	اسمی
قیمت	پیوسته	عددی

جدول 1

3. نمودار های مناسب داده ها

داده	نمودار مناسب
سال ساخت	Box plot
مساحت	Histogram, Box plot
طبقه	Bar chart, Histogram
تعداد کل طبقات	Bar chart
نوع خانه	Pie chart, Bar chart
تعداد اتاق	Bar chart, Histogram
آسانسور دارد	Pie chart, Bar chart
پارکینگ دارد	Pie chart, Bar chart
سطح امنیت منطقه	Bar chart
نوع پوشش کف واحد	Pie chart, Bar chart
قیمت	Histogram, Box plot

جدول 2

دلیل نمودار های انتخاب شده:

1. **Histogrsm: این نمودار برای داده های عددی مانند مساحت، قیمت و ... مناسب است تا توزیع داده ها را نشان دهد.**
2. **Box plot: این نمودار برای نمایش پراکندگی داده های عددی مانند مساحت و قیمت مناسب است.**
3. **Bar chart: این نمودار برای نشان دادن ویژگی های گسسته و دسته بندی شده مانند تعداد اتاق، طبقه، تعداد کل طبقات، سطح امنیت منطقه و ... مناسب است.**
4. **Pie chart: از این نمودار برای ویژگی های اسمی و دودویی مانند وجود یا عدم وجود آسانسور، پارکینگ، نوع خانه و ... استفاده میشود تا نسبت هر دسته را نمایش دهد.**

4. نمودار مناسب جهت نشان دادن ارتباط بین مساحت خانه و قیمت

به نظر من مناسب ترین نمودار جهت نشان دادن ارتباط بین مساحت خانه ها و قیمت آن ها، نمودار پراکندگی یا همان scatter plot است چرا که این نمودار نشان می دهد چگونه قیمت خانه با تغییرات مساحت آن تغییر می کند و آیا رابطه ای مانند همبستگی مثبت و یا منفی بین این دو فیچر وجود دارد یا خیر.

5. میزان همبستگی ارتباط بین مساحت خانه و قیمت

به نظر من در این سوال میزان همبستگی را باید مثبت در نظر گرفت چرا که به طور کلی و عموما قیمت خانه با بالا رفتن مساحت آن افزایش می یابد ولی با توجه به این که در سوال گفته شده خانه هایی وجود دارند که با مساحت کمتر ممکن است قیمت بالاتری داشته باشند، که این موضوع را مثلا میتواند به دلیل موقعیت جغرافیایی و محله آن خانه در نظر گرفت. به این ترتیب اگر تعداد این استثناها زیاد باشند به طوری که نتوان الگوی مشخصی بین این دو فیچر کشف کرد میتوانیم میزان همبستگی را مثبت و نزدیک به صفر فرض کنیم ولی منطقا تعداد این خانه های استثنا زیاد نیستند پس من فکر میکنم میزان همبستگی مثبت و بیشتر از ۰.۵ می باشد و احتمالا میزان همبستگی به ۱ نزدیک تر است.

سوال ۲

1. محاسبه میانگین، انحراف‌معیار، میانه و چارک‌های اول و سوم

Math = [17.5, 16, 17, 15, 15, 16, 18, 20, 13.5, 17, 16, 18, 13.5, 20, 15, 16.4, 16.5, 15, 17, 18]

Physics = [20, 81, 80, 70, 85, 75, 87, 97, 35, 90, 83, 88, 69, 100, 81, 10, 79, 78, 84, 91]

میانگین:

$$\overline{x} = \frac{1}{n}\sum_{1}^{n} x_i$$

میانگین نمرات ریاضی = ۱۶.۵۲۵

میانگین نمرات فیزیک = ۷۴.۱۵

انحراف معیار:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1}}$$

انحراف معیار نمرات ریاضی = ۱.۷۸۷

انحراف معیار نمرات فیزیک = ۲۴.۲۴۲

میانه:

مقادیر مرتب شده برای دروس:

Math: [13.5, 13.5, 15, 15, 15, 15, 16, 16, 16, 16.5, 16.5, 17, 17, 17, 17.5, 18, 18, 18, 20, 20]

Physics: [10, 20, 35, 69, 70, 75, 78, 79, 80, 81, 81, 83, 84, 85, 87, 88, 90, 91, 97, 100]

$$Math\ Median = \frac{16.5 + 16.5}{2} = 16.5$$

$$Physics\ Median = \frac{81 + 81}{2} = 81$$

چارک‌های اول و سوم:

نمرات ریاضی:

به دست آوردن مکان چارک اول: $Q_1 = \frac{25}{100} \times (20 + 1) = 5.25$

که این مقدار نشان میدهد اگر داده ها را مرتب کنیم، چارک اول بین مقادیر پنجم و ششم به دست می آید که در این صورت چارک اول برابر است با:

$$Q_1 = \frac{15 + 15}{2} = 15$$

به همین ترتیب برای محاسبه چارک سوم داریم:

به دست آوردن مکان چارک سوم: $Q_3 = \frac{75}{100} \times (20 + 1) = 15.75$

که این مقدار نشان میدهد اگر داده ها را مرتب کنیم، چارک سوم بین مقادیر پانزدهم و شانزدهم به دست می آید که در این صورت چارک سوم برابر است با:

$$Q_1 = \frac{17.5 + 18}{2} = 17.625$$

نمرات فیزیک:

به دست آوردن مکان چارک اول: $Q_1 = \frac{25}{100} \times (20 + 1) = 5.25$

که این مقدار نشان میدهد اگر داده ها را مرتب کنیم، چارک اول بین مقادیر پنجم و ششم به دست می آید که در این صورت چارک اول برابر است با:

$$Q_1 = \frac{70 + 75}{2} = 72.5$$

به همین ترتیب برای محاسبه چارک سوم داریم:

به دست آوردن مکان چارک سوم: $Q_3 = \frac{75}{100} \times (20 + 1) = 15.75$

که این مقدار نشان میدهد اگر داده ها را مرتب کنیم، چارک سوم بین مقادیر پانزدهم و شانزدهم به دست می آید که در این صورت چارک سوم برابر است با:

$$Q_1 = \frac{87 + 88}{2} = 87.5$$

2. ارزیبای عملکرد کل کلاس بر اساس این معیار ها

میانگین:

این معیار نمایانگر مقدار متوسط نمرات کل کلاس است و برای مجموعه داده هایی که توزیع یکنواخت دارند معیار مناسبی است. اما این معیار به شدت تحت تاثیر ناهنجاری ها قرار می گیرد و اگر چند نمره بسیار کم و یا چند نمره بسیار زیاد در کلاس وجود داشته باشد، این ناهنجاری ها میانگین مقدار واقعی عملکرد کلاس را منحرف می کنند.

انحراف معیار:

این معیار نشان دهنده میزان پراکندگی نمرات کلاس است و اگر مقدار آن کم باشد یعنی نمرات نزدیک به هم هستند و اگر مقدار آن زیاد باشد نشان دهنده عدم یکنواختی در سطح کلاس است. این معیار مانند میانگین به ناهنجاری ها حساس است و عدد به دست آمده فقط مقدار پراکندگی را نشان می دهد و اطلاعات دقیقی درباره مرکز توزیع نمرات ارائه نمی دهد. به همین منظور به نظر من انحراف معیار می تواند به عنوان شاخصی برای بررسی یکنواختی کلاس مفید باشد اما برای بررسی عملکرد کل کلاس کافی نیست.

میانه:

این معیار به ناهنجاری ها حساس نیست و حتی اگر یک دانش آموز نمره خیلی کم و یا خیلی زیادی گرفته باشد، روی مقدار میانه تاثیر زیادی ندارد. اگر داده ها به شدت متقارن باشد، میانگین و میانه تقریبا برابر خواهند بود و تفاوتی ایجاد نمی شود ضمن این که این معیار صرفا یک مقدار را نمایش می دهد و پراکندگی را نشان نمی دهد.

چارک اول و سوم:

چارک ها اطلاعات بیشتری از پراکندگی نمرات را ارائه می دهند. محدوده بین چارک اول و سوم نشان می دهد که ۵۰ درصد میانی دانش آموزان چگونه عمل کرده اند. از این معیار می توان برای شناسایی ناهنجاری ها و میزان یکنواختی نمرات در کلاس استفاده کرد.

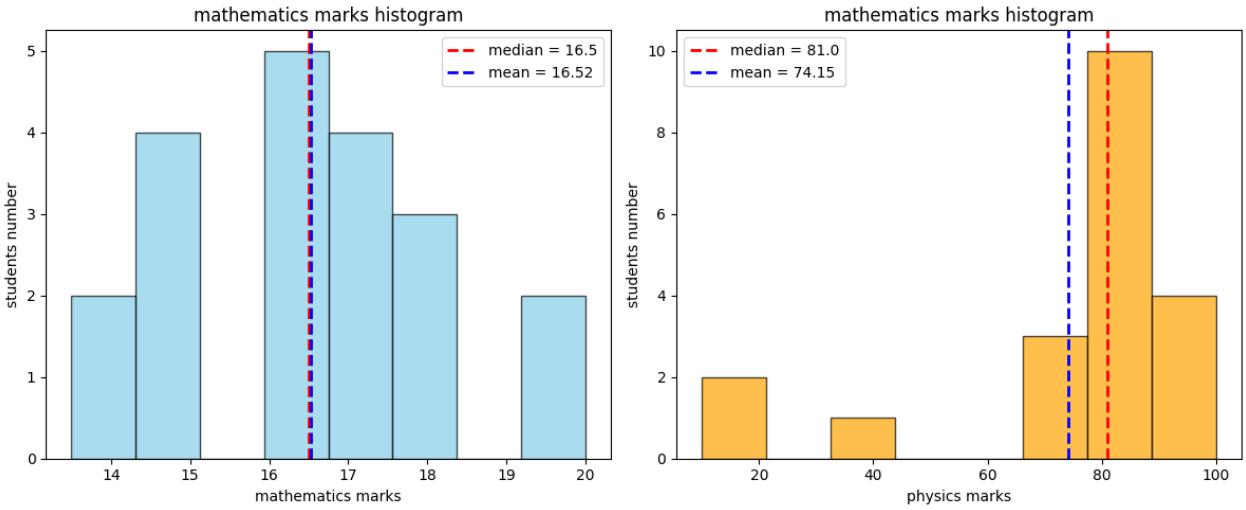
جمع‌بندی:

به نظر من برای نمراتی مثل فیزیک که دارای ناهنجاری هست استفاده از میانه مناسب تر است در مقابل برای داده های درس ریاضی که دارای ناهنجاری کمتری هستند، استفاده از میانگین بهتر است. همچنین اگر بخواهیم یکنواختی عملکرد کلاس را بررسی کنیم به نظرم استفاده از انحراف معیار بهتر است و اگر هم بخواهیم بررسی دقیق تری روی توزیع نمرات داشته باشیم میتوانیم از چارک ها به همراه میانه استفاده کنیم.

3. بررسی ناهنجاری در داده ها

با توجه به داده ها و بررسی چارک های اول و سوم میتوان اینطور برداشت کرد که نمرات ریاضی ناهنجاری زیادی ندارند و این داده ها به طور کلی دارای یکنواختی هستند چرا که نمرات داده شده فاصله چندانی با چارک های اول و سوم ندارند. در حالی که نمرات فیزیک دارای ناهنجاری زیادی هستند برای مثال نمره های ۱۰، ۲۰، ۳۵ نمراتی هستند که به عنوان ناهنجاری در نظر گرفت که می تواند عملکرد کلی کلاس را تحت تاثیر قرار می دهند.

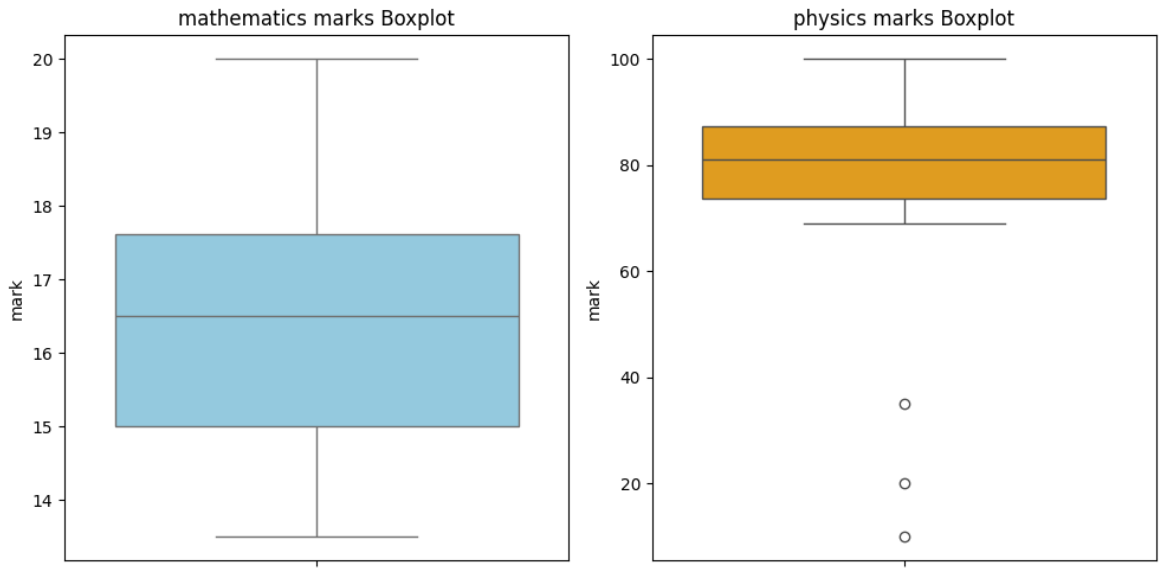
4. رسم نمودار هیستوگرام



نمودار 1

همانطور که از نمودار ها مشخص است، توزیع نمرات ریاضی به مراتب یکنواخت تر بوده و میانگین و میانه تقریبا با هم برابر هستند و این موضوع به دلیل متقارن بودن نمرات است. همچنین میتوان گفت در این توزیع داده ناهنجار هم وجود ندارد. در حالی که در نمودار نمرات درس فیزیک، توزیع یکنواخت نیست و میانگین و میانه هم باهم اختلاف دارند که به دلیل عدم تقارن در داده ها است. همچنین این توزیع دارای داده های ناهنجار هم می باشد که با متوسط نمرات اختلاف فاحشی دارند.

5. رسم نمودار Boxplot



نمودار 2

همانطور که از نمودار های رسم شده مشخص است، نمرات ریاضی دارای توزیع متعادلی هستند و داده های این توزیع بدون ناهنجاری اند. در این توزیع طبق نمودار رسم شده، میانه نزدیک به مرکز جعبه است که نشان می دهد توزیع تقریبا متقارن است.

در حالی که طبق نمودار رسم شده برای نمرات درس فیزیک می توان برداشت کرد که توزیع نمرات چولگی منفی دارد به ای معنا که تعداد کمی از دانش آموزان نمرات خیلی پایینی دارند اما اکثریت در بازه ۸۰ تا ۱۰۰ هستند. در این توزیع ناهنجاری ها تاثیر زیادی بر میانگین دارند و به نظر من در این درس میانه می تواند معیاری بهتری نسبت به میانگین برای بررسی عملکرد کل درس باشد.

6. نرمال سازی توزیع ها

من برای این قسمت، نرمال سازی را با استفاده از روش Z-score انجام دادم که با استفاده از کد پایتونی که در ادامه آورده شده است داده های نمرات فیزیک و ریاضی را نرمال سازی کرده ام. در این نرمال سازی میانگین داده ها از مقدار اصلی داده کم شده و حاصل بر انحراف معیار داده ها تقسیم می شود که جواب حاصل شده مقدار نرمال شده آن داده است. این روش داده ها را مقیاس بندی می کند و تاثیر واحدهای اندازه گیری مختلف را کاهش می دهد. همچنین باعث می شود که داده ها دارای میانگین صفر و انحراف معیار ۱ شوند.

کد پایتون استفاده شده:

```
math_scores = np.array([13.5, 13.5, 15, 15, 15, 15, 16, 16, 16, 16.5, 16.5, 17,
17, 17, 17.5, 18, 18, 18, 20, 20])
physics_scores = np.array([10, 20, 35, 69, 70, 75, 78, 79, 80, 81, 81, 83, 84,
85, 87, 88, 90, 91, 97, 100])

math_mean = np.mean(math_scores)
math_std = np.std(math_scores)

physics_mean = np.mean(physics_scores)
physics_std = np.std(physics_scores)
math_z = (math_scores - math_mean) / math_std
physics_z = (physics_scores - physics_mean) / physics_std
print(math_z)
print(physics_z)
```

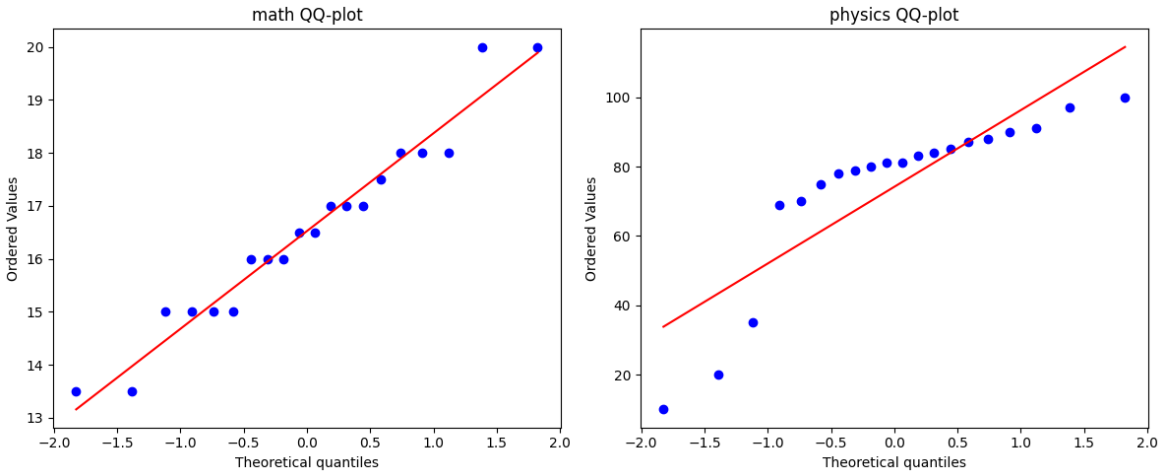
توزیع های نرمال شده:

Math: [-1.74, -1.74, -0.88, -0.88, -0.88, -0.88, -0.3 , -0.3 , -0.3 , -0.01, -0.01 , 0.27 , 0.27 , 0.27 , 0.27 , 0.56 , 0.85, 0.85, 0.85, 1.99, 1.99]

Physics = [-2.71, -2.29, -1.66, -0.22, -0.18, 0.04, 0.16, 0.21, 0.25, 0.29, 0.29, 0.37, 0.42, 0.46, 0.54, 0.59, 0.67, 0.71, 0.97, 1.09]

7. بررسی نمودار QQ – plot

نمودار QQ-plot داده های مشاهده شده را با یک توزیع نرمال تئوری مقایسه می کند، اگر داده ها نرمال باشند، نقاط در نمودار روی یک خط صاف قرار میگیرند و اگر انحراف از خط وجود داشته باشد، نشان دهنده آن است که داده ها دارای چولگی و یا کشیدگی غیر طبیعی هستند.



نمودار 3

همانطور که از نمودار های رسم شده مشخص است، در نمودار رسم شده برای نمرات ریاضی داده ها تقریبا روی خط قرمز قرار دارند که نشان می دهد نمرات ریاضی نزدیک به نرمال هستند در صورتی که نمودار نمرات فیزیک نشان میدهد در بخش های ابتدایی و انتهایی نقاط از خط قرمز فاصله دارند که نشان دهنده آن است که داده های فیزک دارای چولگی هستند و به طور کامل نرمال نیستند.

8. محاسبه همبستگی بین دو توزیع

من برای این قسمت فرمول محاسبه همبستگی در اسلاید ها رو پیدا نکردم به همین دلیل از فرمول همبستگی پیرسون که در کتاب مرجع بود استفاده کردم که این فرمول به شرح زیر است:

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{166.79}{454.87} = 0.367$$

از آنجایی که همبستگی مثبت است اما به ۱ نزدیک نیست، این نشان دهنده آن است که همبستگی مثبت ضعیف بین نمرات ریاضی و فیزیک وجود دارد. به همین دلیل نمیتوان با اطمینان گفت که دانش آموزانی که در ریاضی نمره بالاتری دارند، حتما در فیزیک هم عملکرد خوبی داشته اند. این موضوع می تواند به دلیل وجود داده های پرت در درس فیزیک و یا کم بودن داده های دو درس باشد.

9. بررسی داده های پاک سازی شده

برای داده های ریاضی داریم:

N/A: دلیل حذف این داده این است که احتمالا دانش آموز در آن جلسه غایب بوده و یا به هر دلیل نمره ای برای درس ریاضی برای او ثبت نشده است پس برای تاثیر نگذاشتن این داده روی داده های دیگر آن را حذف کرده اند.

21: این داده از سقف مجاز نمرات برای این درس بالاتر بوده است که به همین دلیل آن را حذف کرده اند. این داده می تواند به دلیل این باشد که آن دانش آموز نمرات امتیازی بیشتری در این درس داشته و به همین دلیل این نمره را کسب کرده و به نظر من بهتر بود آن را با نمره ۲۰ جایگزین کنند.

0: این نمره احتمالا به این دلیل حذف شده که نه تنها با میانگین نمرات، بلکه با کمینه نمرات نیز فاصله بسیار زیادی دارد و به نظر من به عنوان داده پرت شناسایی شده است.

+A: این نمره خارج از الگوی نمره دهی برای این درس است و به همین منظور حذف گردیده است.

”19“: این نمره دارای فرمت نوشتاری اشتباه است و به همین دلیل از نمرات درس حذف شده است. وجود علامت کوتیشن برای این نمره اشتباه نوشتاری است و به نظر من راه بهتر آن بود که با استفاده از ابزارهایی مانند عبارات منظم در پایتون مقدار اصلی نمره را استخراج کنند.

برای درس فیزیک داریم:

90, 76: دلیل حذف این نمرات برای من مشخص نیست و به نظرم بهتر بود این نمرات حذف نشوند چرا که مشکلی ندارند. با توجه به نمرات درس فیزیک این نمرات جزو داده های پرت حساب نمی شوند و موردی هم برای قرارگیری در دیتاست ندارند. شاید منظور نویسنده برای حذف این نمرات به دلیل هم اندازه کرده تعداد داده ها در دو دیتا ست دروس ریاضی و فیزیک باشد.

0: دلیل حذف این نمره مانند درس ریاضی می تواند به دلیل این باشد که این داده جزو داده های پرت محسوب می شود و با میانگین و کمینه نمرات هم فاصله بسیار زیادی دارد.

B: همانطور که در بخش قبل گفته شد این نمره خارج از فرمت و الگوی نمره دهی برای این درس است پس باید حذف شود.

"84": این نمره هم مانند بخش قبل دارای فرمت نوشتاری اشتباه است و وجود علامت کوتیشن می تواند دلیل حذف این نمره باشد اگر چه به نظر من بهتر بود که نمره را از این فرمت استخراج نمایند.

سوال ۳

برای این سوال جدول فراوانی و جدول مقادیر مورد انتظار به شرح زیر است:

جنسیت / رشته	کامپیوتر	برق	مکانیک	مجموع
پسر	40 (38.18)	50 (54.55)	30 (27.27)	۱۲۰
دختر	30 (31.82)	50 (45.45)	20 (22.73)	۱۰۰
مجموع	۷۰	۱۰۰	۵۰	۲۲۰

جدول 3

فرض صفر (H_0) : انتخاب رشته مستقل از جنسیت است.

فرض جایگزین (H_A) : انتخاب رشته وابسته به جنسیت است.

برای محاسبه آماره کای – ۲ داریم:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

که این مقدار برابر است با:

$$\chi^2 = 1.62$$

همچنین درجه آزادی برابر است با:

$$df = (2-1) \times (3-1) = 2$$

در ادامه باید از تابع توزیع تجمعی کای – ۲ استفاده کنیم و مقدارp-value را محاسبه کنیم که داریم:

$$p = 1 - F_{\chi^2}(1.62, 2) = 0.4448$$

که طبق محاسبات انجام شده مقدار *p-value* برابر *0.4448* و بزرگتر از *significance level* است. بنا بر این تفاوت بین مقدار مشاهده شده و مقدار مورد انتظار تصادفی است و داده های ما شواهد کافی برای رد فرض صفر را ارائه نمی دهند و فرض صفر *(H0)* را رد نمی کنیم. یعنی جنسیت و انتخاب رشته مستقل از هم هستند.

سوال عملی

2. بارگزاری و نمایش داده ها

برای این قسمت چون کد در محیط کولب نوشته شده، ابتدا به گوگول درایو متصل شدم و سپس با استفاده از کتابخانه pandas فایل های csv داده ها را خواندم و آن ها را در دیتا فریم مربوطه ریختم و با توجه به خواسته سوال برای هر شهر ۵ نمونه از داده هایش را چاپ کردم که نتایج در فایل نوت بوک قابل مشاهده است.

3. تجمیع داده ها

در این مرحله سعی کردم داده ها را تجمیع کنم که در این راه به چند مشکل برخوردم که در ادامه هر مشکل و راه حل مربوط به آن را به طور کامل شرح می دهم.

ابتدا اولین مشکل این بود که تعداد ستون های دیتا فریم شهرهای مختلف با هم برابر نبود، شهر مشهد ستون قیمت بر حسب متر مربع برای هر خانه را نداشت و شهر اصفهان هم ستون exchangeable را به عنوان ستون اضافی داشت. با توجه به این که این ستون فقط و فقط در داده های شهر اصفهان موجود بود، این ستون را از دیتا فریم داده های اصفهان حذف کردم و ستون قیمت بر حسب متر مربع را هم به داده های شهر مشهد اضافه کردم. با توجه به این که مقادیر ستون قیمت بر حسب متر مربع در دیتافریم شهر مشهد بر اساس قیمت کل و متراژ خانه به دست می آمد، فعلا مقادیر این ستون را برابر NaN قرار دادم تا در ادامه بعد از تمیز کردن داده ها این ویژگی را محاسبه کنم.

در مرحله بعد نام ستون های شهر های مختلف را تغییر دادم و به یک شکل تبدیل کردم تا بتوانم با استفاده از متد concat در کتابخانه pandas دیتا فریم های مختلف را در یک دیتا فریم تجمیع کنم.

4. شناسایی و حذف داده های نادرست

برای این مرحله من متوجه شدم آگهی هایی که به درستی کرال شدند، همگی دارای الگوی مشخصی هستند و آگهی های درست در قسمت قیمت کل خانه ها دارای واژه ”تومان” می باشند.

به همین دلیل با استفاده دیتا فریم clean_df را تعریف کردم و همه آگهی هایی که در ستون total_priceشان دارای واژه تومان بودند را در این دیتا فریم ریختم که این ها آگهی هایی هستند که به طور صحیح کرال شده اند. تعداد آگهی های قبل از این تغییر برابر ۹۴۵۱ آگهی و تعداد آگهی ها بعد از این پاکسازی برابر ۵۹۳۲ آگهی بوده است.

5. بررسی آگهی های تکراری

برای این قسمت الگویی که من برای شناسایی آگهی های تکراری پیدا کردم این بود که در این آگهی ها، مقادیر یه ستون title, property_size, total_price مقادیر یکسانی بوده اند و آگهی هایی که دارای مقادیر یکسان در این سه ستون هستند، آگهی های تکراری هستند. به همین دلیل با استفاده از متد duplicated در کتابخانه pandas آگهی هایی که در این سه ستون مقادیر مشابه داشتند را در دیتا فریم df_duplicates ریختم و آن ها را به صورت مرتب شده چاپ کردم که برای شما قابل مشاهده باشد. تعداد این آگهی ها برابر با ۲۴۳۹ عدد آگهی بود. سپس با استفاده از متدdrop_duplicates این آگهی های تکراری را حذف کردم و فقط اولین آگهی از آگهی های تکراری را در دیتا فریم نگه داشتم.

6. بررسی نوع داده ها

با استفاده از دستور dtype() نوع داده های هر ستون را چاپ کردم که نوع داده هر ستون و نوع داده ای که باید به آن تبدیل شود را در جدول زیر آورده ام:

داده	نوع داده قبل از تغییر	حجم قبل از تغییر	نوع داده بعد از تغییر	حجم بعد از تغییر
title	object	1215.77	object	1215.77
Property_size	object	350.05	Int 64	49.55
Total_price	object	816.71	Int 64	49.55
Price_per_meter	object	627.75	Int 64	49.55
Room_count	object	320.76	Int 64	49.55
Build_year	object	337.63	Int 64	49.55
Floor_count	object	340.25	Int 64	49.55
Total_floor	object	340.26	object	49.55
characteristics	object	1582.01	object	1582.01
features	object	2287	object	2287
description	object	8189.58	object	8189.58
url	object	1674.60	object	1674.60
Crawl_date	object	368.83	Datetime64	44.04
city	object	348.71	object	348.71

جدول 4

برای اینکه بتوانم نوع داده ها را اصلاح کنم ابتدا با استفاده از regex برای داده های عددی صرفا مقادیر عددی آن ها را ذخیره کردم و عبارت های غیر عددی مانند "تومان" را از بین این داده ها پام کردم همچنین مقادیر Not found را هم با مقدار NaN خود کتابخانه pandas جایگذاری کردم تا بتوانم فرمت داده ها را به Int64 تغییر دهم. همچنین بعد از اینکه فرمت داده ها را اصلاح کردم برای داده های ستون price_per_meter که ممکن بود مقادیر غیر قابل قبول وجود داشته باشد و یا در قسمتی که این ستون را به شهر مشهد اضافه کرده بودیم مقادیر NaN دریافت کرده باشند را با استفاده از تقسیم مقادیر ستون total_price بر property_size محاسبه کردم و این مقادیر گمشده را جایگزین کردم.

7. پردازش ستون های ویژگی ها و امکانات

در این قسمت سعی کردم ستون های ویژگی ها و امکانات را بررسی کنم و اگر داده با اهمیتی در آن ها بود، آن را استخراج کنم. برای این کار ابتدا ۳۰ نمونه اول ستون characteritics را چاپ کردم و الگویی که در این ستون قابل مشاهده بود، وضعیت بازسازی خانه ها بوده است. سپس همین کار را برای ستون features هم تکرار کردم. برای این که متوجه شوم کدام عبارات در این ستون ها پر تکرار هستند و بتوانم از این ستون ها ویژگی هایی را استخراج کنم، ابتدا مقادیر NaN را حذف کردم و سپس با توجه به این که این ویژگی ها با عبارت "|" از هم جدا شده بودند، آن ها را جدا کردم و مقادیر در این ستون ها را به شکل لیست درآوردم و دیتا فریم clean_df2 را تشکیل دادم. در مرحله بعد یا استفاده از متد unique در کتابخانه pandas درصد تکرار هر کدام از این ویژگی ها را استخراج کردم که متوجه شدم ویژگی های زیر بیشترین تکرار را دارند:

پارکینگ: ۱۰.۰۹ ٪

انباری: ۹.۹۱ ٪

آسانسور: ۹.۲۳ ٪

وضعیت بازسازی واحد: ۹.۱۵ ٪

در نتیجه در مرحله بعدی تابع هایی را تعریف کردم که این ویژگی ها را در دیتافریم شناسایی کرده و در صورت وجود هر کدام از این امکانات در ستون مربوطه که تازه تعریف شده مقدار ۱ قرار دهد و اگر آن ساختمان آن ویژگی را نداشت ۰ قرار دهد.

8. بررسی محتوای ستون توضیحات و عنوان

در این قسمت با توجه به گفته سوال، ۱۰ نمونه ستون های عنوان و توضیحات را به صورت تصادفی چاپ کردم. طبق مطالب چاپ شده ممکن است عنوان و یا توضیحات اطلاعاتی را به ما بدهد مانند اینکه وضعیت جغرافیایی واحد چگونه است؟ و یا این که وضعیت نورگیر بودن خانه به چه صورت است؟ و ...

اما به طور کلی این دو دسته داده بسیار به هم ریخته و نامرتب هستند و جدای از این موضوع موارد این دو ستون برای همه واحد ها یکسان نیستند. یعنی برای مثال ممکن است در یک آگهی موقعیت جغرافیایی خانه شرح داده شده باشد اما در آگهی های دیگر این موقعیت توضیح داده نشده باشد و نتوانیم از این اطلاعات به عنوان ویژگی متمایز کننده خانه ها استفاده کنیم.

9. پردازش داده های گمشده

در این قسمت بررسی کردم که چه داده هایی از آگهی های داده شده، مقادیر گمشده هستند و تعداد مقادیر گمشده هر ستون را استخراج کردم که به شرح زیر است:

داده	تعداد مقادیر گمشده
Property_size	72
Price_per_meter	39
Room_count	82
Build_year	76
Floor_count	2671
Total_floor	2666
Crawl_Date	74

جدول 5

برای یافتن مقادیر گمشده ستون property_size *به* این صورت عمل کردم که به ازای داده هایی که property_size برای آن ها مقدار NaN داشته ولی مقدار *total_price, price_per_meter* آن ها مقدار گمشده نبوده است، متراژ خانه را از طریق تقسیم قیمت کل بر قیمت متر مربع آن خانه محاسبه کردم و در ستون *property_size* مربوطه قرار دادم.

در ادامه برای ستون *room_count* به این صورت عمل کردم که متوسط متراژ خانه ها با اتاق های مختلف را محاسبه کردم. برای مثال متراژ متوسط خانه هایی با ۱ اتاق، متراژ متوسط خانه هایی با ۲ اتاق و ... که نتایج به شرح زیر است:

تعداد اتاق	متوسط متراژ خانه
1	60.99
2	105.73
3	165.54
4	353.93

جدول 6

سپس برای خانه هایی که دارای مقادیر گمشده در ستون تعداد اتاق بودند، متراژ آن خانه را با متوسط متراژ بر حسب تعداد اتاق که در بالا مشاهده می شود، مقایسه کردم و با توجه به نزدیکی متراژ خانه به متوسط متراژ محاسبه به ازای تعداد اتاق، تعداد اتاق ها برای این خانه ها را در دیتا فریم قرار دادم.

در مرحله بعد برای جایگزین کردن داده های گم شده در ستون سال ساخت خانه، با توجه به این که تعداد این داده ها زیاد نبود، آن ها را با میانگین رند شده سال ساخت همه خانه ها جایگزین کردم.

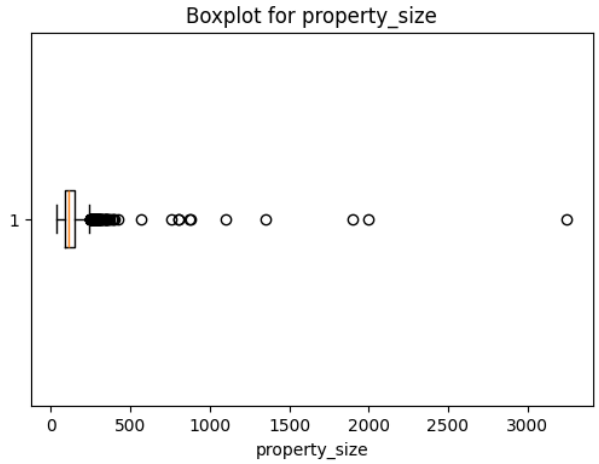
سپس برای جایگزینی مقادیر گمشده برای تعداد طبقات و طبقه واحد ابتدا میخواستم آن ها را با میانه تعداد کل طبقات و میانه تعداد طبقات واحد جایگزین کنم که بعد از این که این کار را انجام دادم، تعداد داده های پرت برای این دو مقدار بسیار زیاد بود که به نظرم دلیل آن می تواند این باشد که برخی از خانه ها کلنگی هستند و اساسا تعداد طبقه برای این خانه ها لحاظ نمی شوند. به همین منظور اینگونه عمل کردم که برای خانه هایی که تنها یک مقدار از دو ستون *floor_count , total_floor* مقدار *NaN* داشته اند، مقدار معتبر یکی از این دو ستون را برای دیگری هم قرار دادم. با اینکار تعداد داده های پرت برای این دو ستون بسیار کاهش یافت.

در نهایت هم برای مقادیر گمشده در ستون *crawl_date*، با توجه به این که تاریخ کرال کردن برای همه آگهی ها یکسان بود، مقدار ثابت *2025-02-15* را قرار دادم.

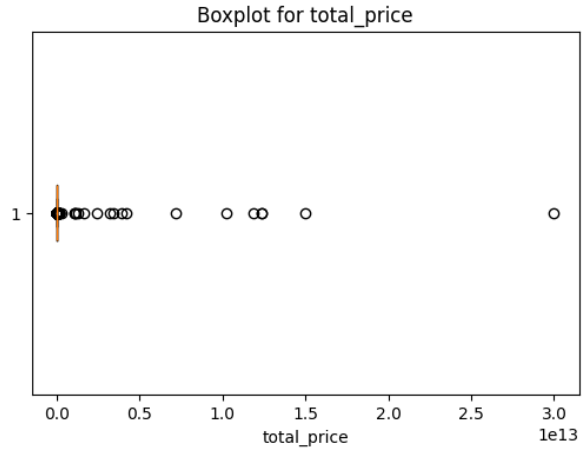
10. شناسایی مقادیر پرت

برای این قسمت و جهت شناسایی مقادیر پرت من از نمودار *Box plot* استفاده کردم و مقادیر پرت برای هر ستون عددی دیتا فریم را رسم کردم.

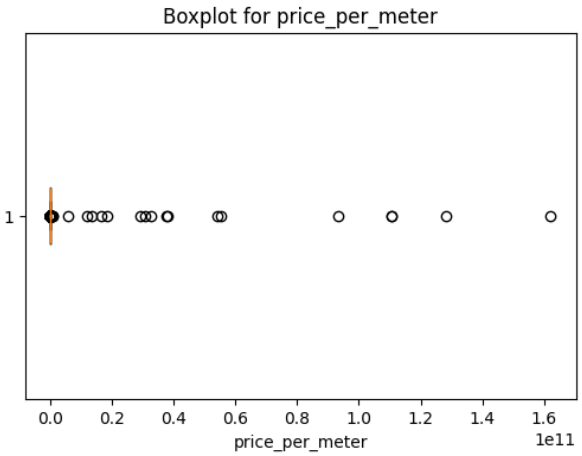
نمودار حاصله برای این مجموعه داده به شرح زیر است:



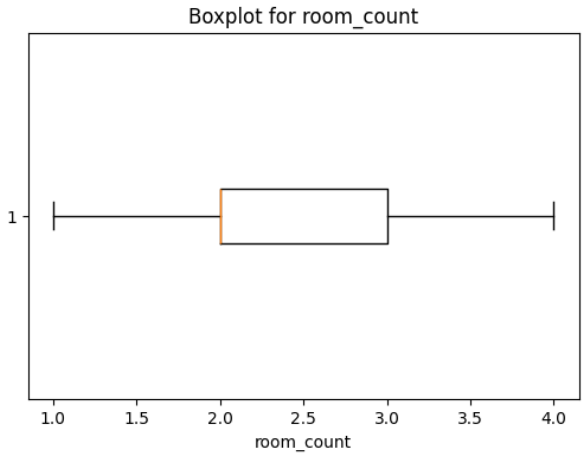
نمودار 4



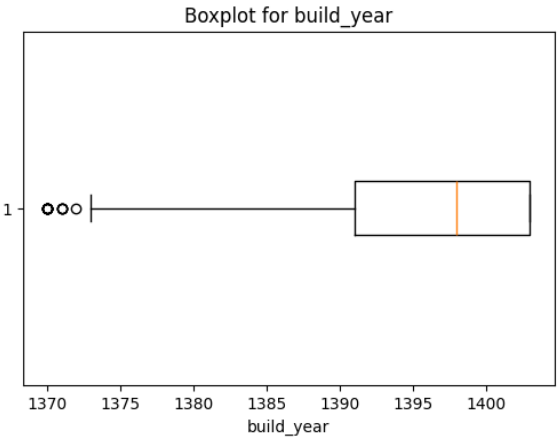
نمودار 5



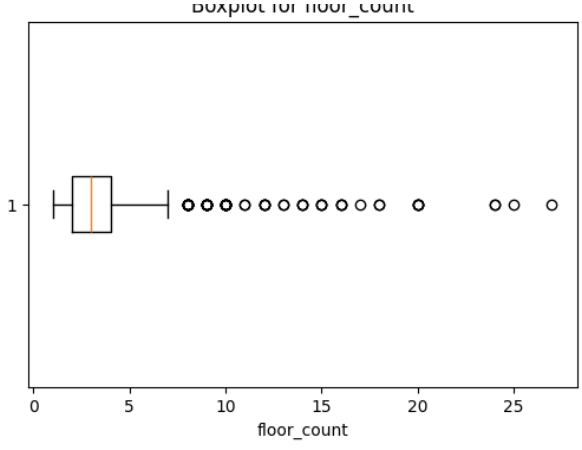
نمودار 6



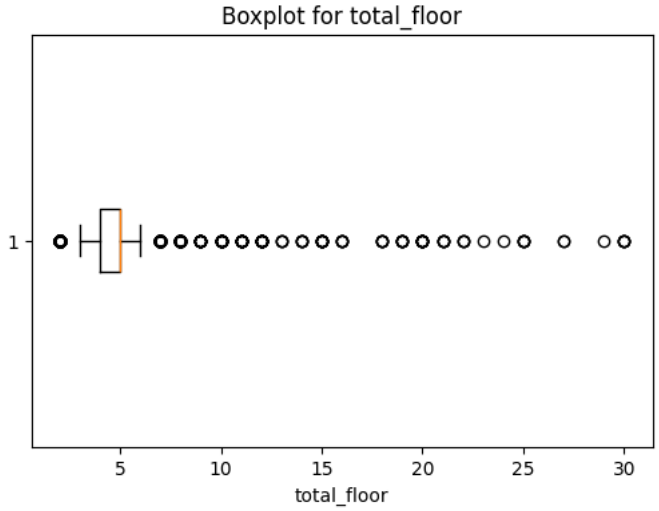
نمودار 7



نمودار 8



نمودار 9



نمودار 10

همچنین برای تعداد مقادیر پرت برای هر ستون داریم:

داده	تعداد مقادیر پرت
Property_size	81
Total_price	277
Price_per_meter	207
Room_count	0
Build_year	17
Floor_count	67
Total_floor	347

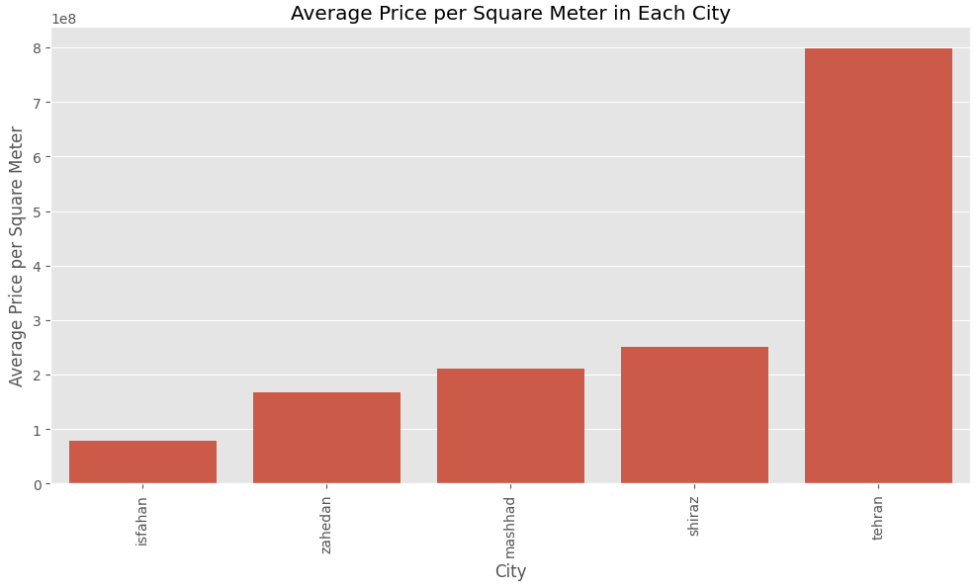
جدول 7

همانطور که از نمودار مشخص است، تعداد داده های پرت برای تعداد کل طبقات بیشترین تعداد را دارد و دلیل آن می تواند عدم وجود داده مناسب و وجود مقادیر گمشده برای این مقدار باشد که در بخش قبلی درباره آن صحبت کردم. برای نمونه برای خانه های کلنگی تعداد کل طبقات بی معنی است و وجود تعداد زیاد مقادیر گمشده در این ستون می تواند دلیلی بر زیاد بودن تعداد مقادیر پرت برای این ستون باشد. در رتبه دوم و سوم تعداد مقادیر پرت که اختلاف خیلی زیاده با بقیه موارد دارند، ستون های قیمت کل و قیمت متر مربع است و دلیل آن می تواند اختلاف قیمت بسیار زیاد بین مناطق مختلف علی الخصوص کلانشهرها باشد.

11. مصور سازی داده ها

برای این بخش من از کتابخانه matplotlib و seaborn استفاده کردم که هر دوی این کتابخانه ها جهت رسم نمودار و مصور سازی داده ها در پایتون هستند.

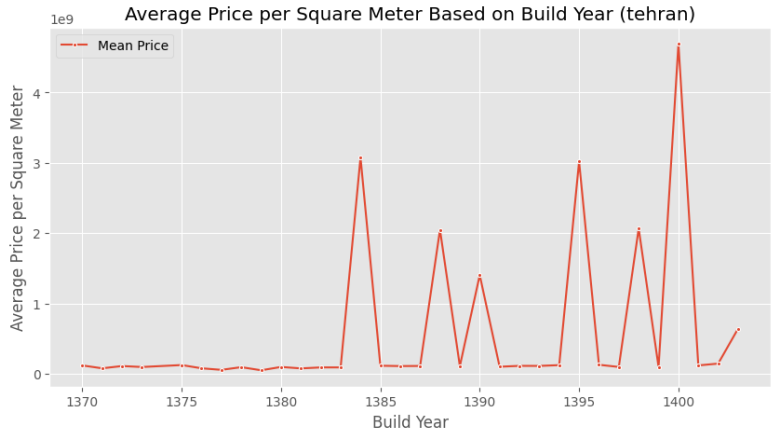
ابتدا کدی را جهت نمایش میانگین قیمت خانه ها بر حسب متر مربع برای هر شهر را نوشتم. فرایند این کد اینگونه است که ابتدا بر اساس شهر داده ها را گروه گروه می کند و سپس برای هر گروه میانگین را محاسبه میکند و این میانگین را طبق نمودار زیر نمایش می دهد:



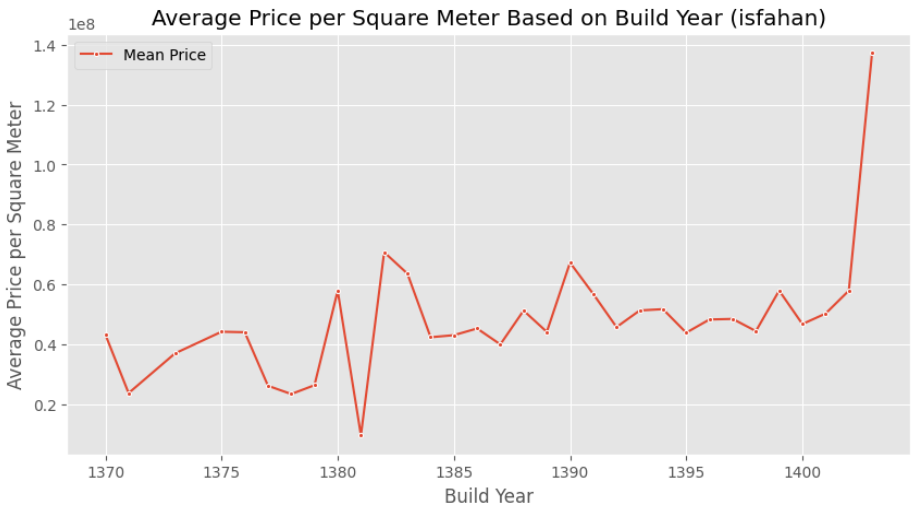
نمودار 11

که همانطور که قابل انتظار است میانگین قیمت بر حسب متر مربع خانه ها در تهران با اختلاف از بقیه شهر ها بیشتر است.

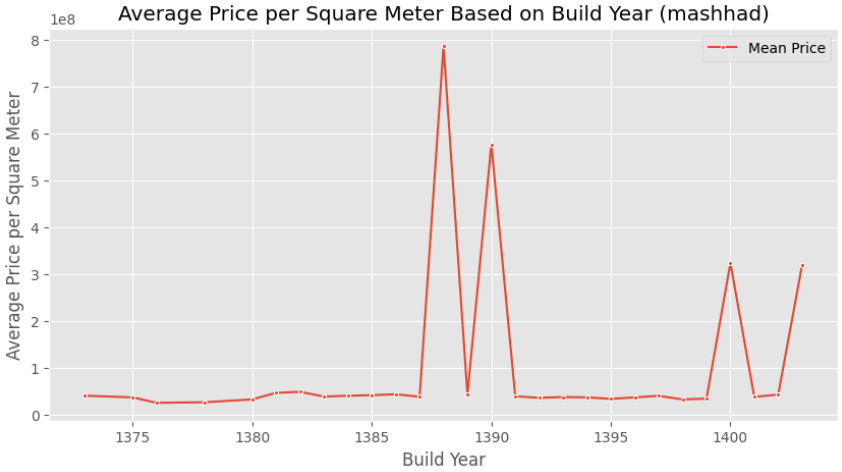
برای قسمت بعد من میخواستم میانگین قیمت بر متر مربع خانه ها را در هر سال نمایش دهم برای همین بجای هیستوگرام از linear plot استفاده کردم که نمودار ها برای هر شهر به شرح زیر است:



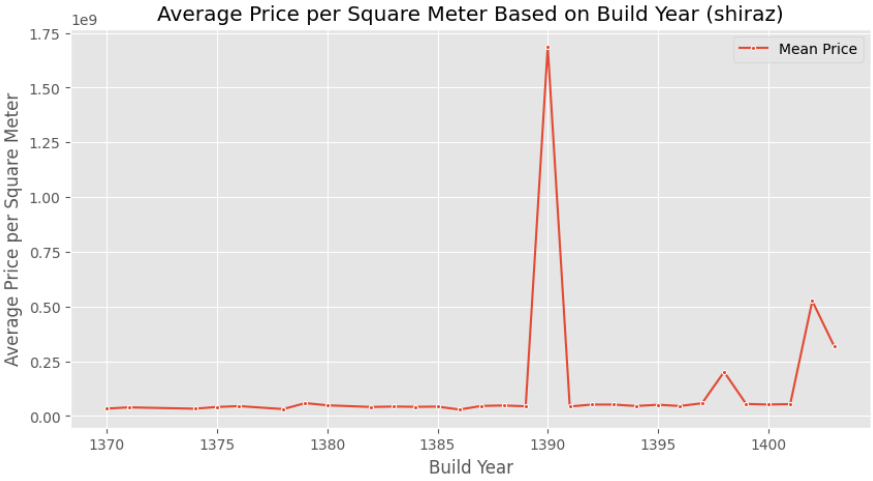
نمودار 12



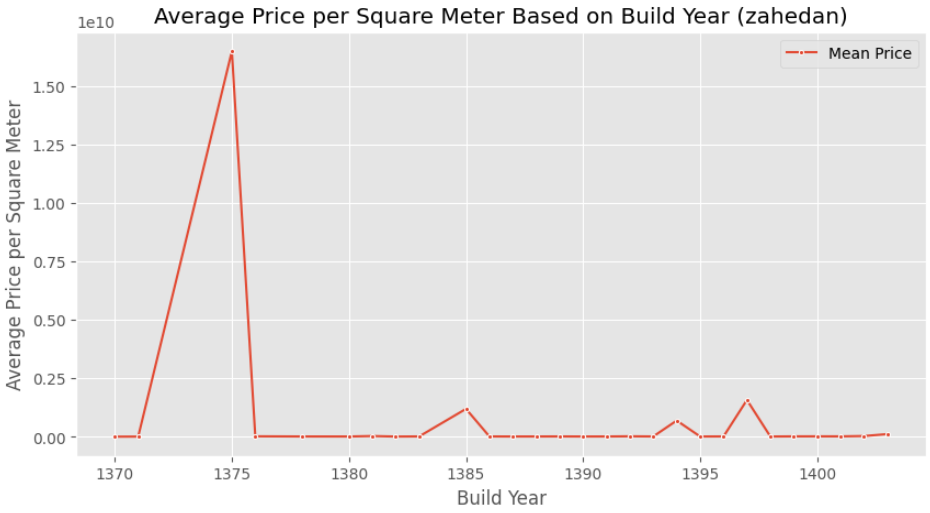
نمودار 13



نمودار 14

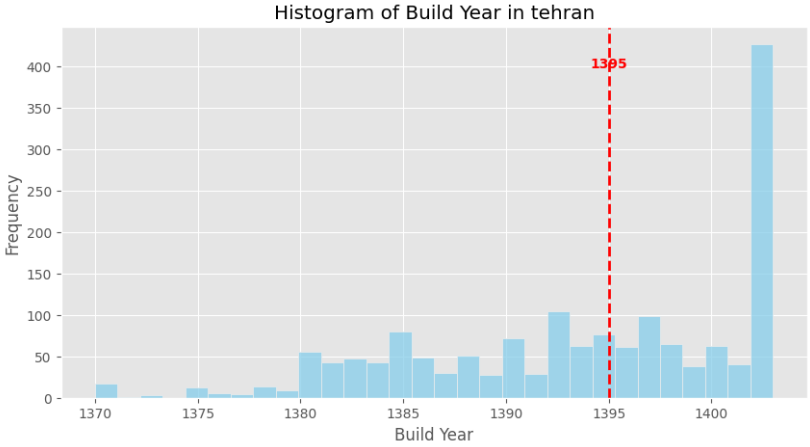


نمودار 15

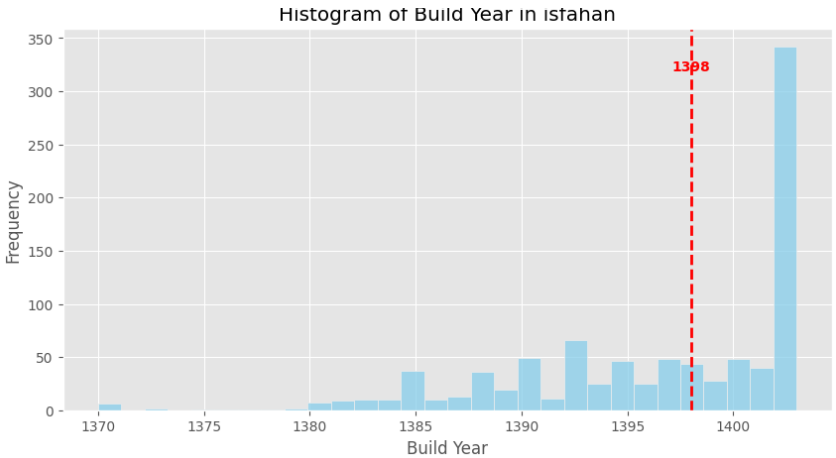


نمودار 16

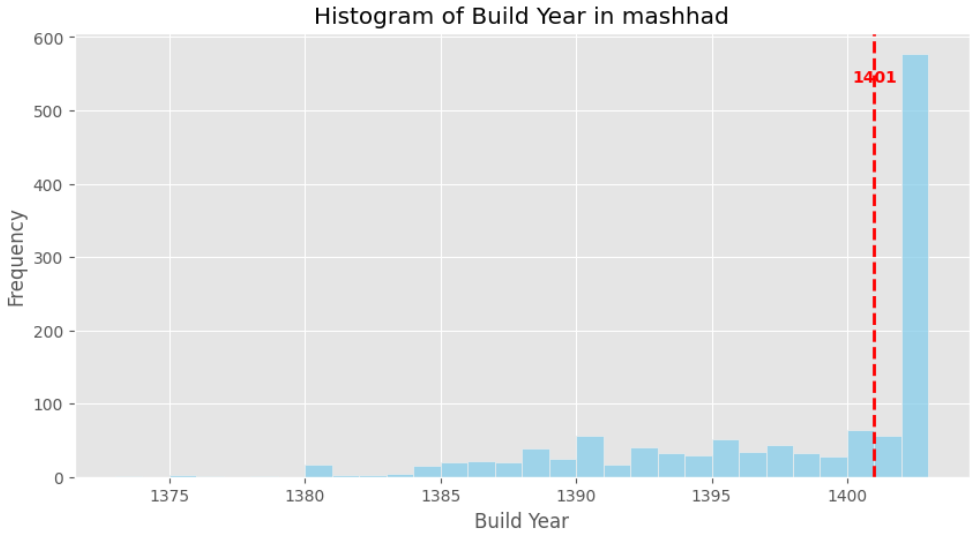
سپس در مرحله بعدی توزیع برای سال ساخت خانه ها را برای هر شهر مصور کردم. در این قسمت نیز مانند قسمت های قبلی با استفاده از گروه کردن شهر های یکسان نمودار داده های آن ها را رسم کردم:



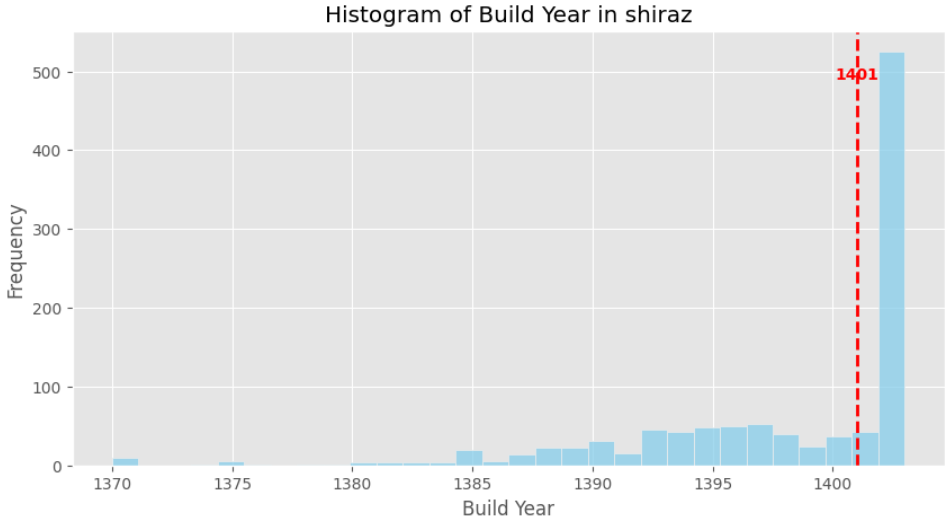
نمودار 17



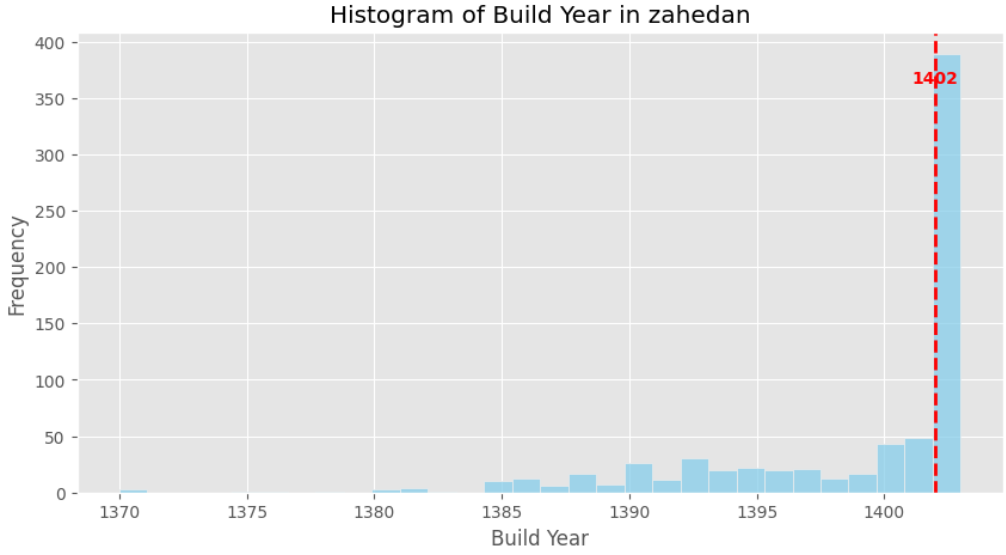
نمودار 18



نمودار 19



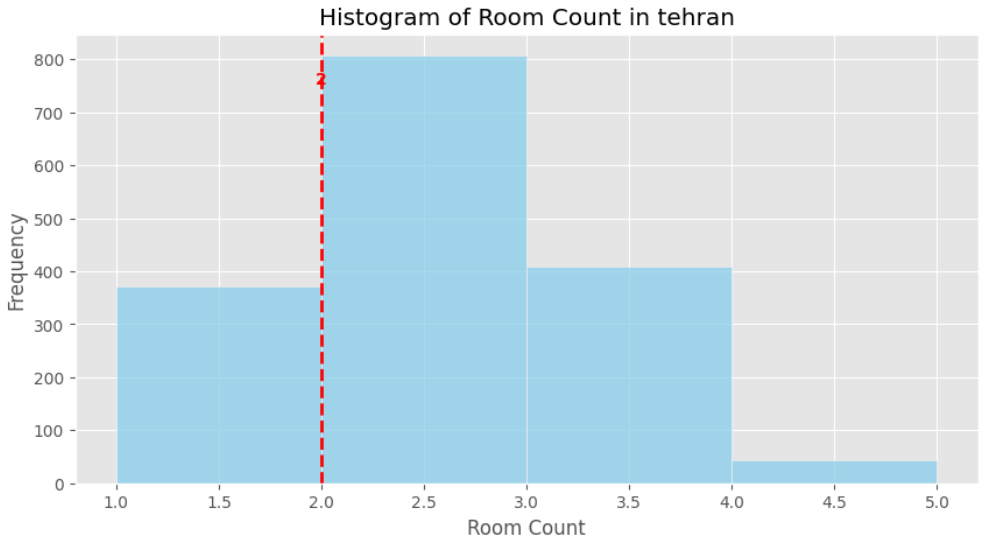
نمودار 20



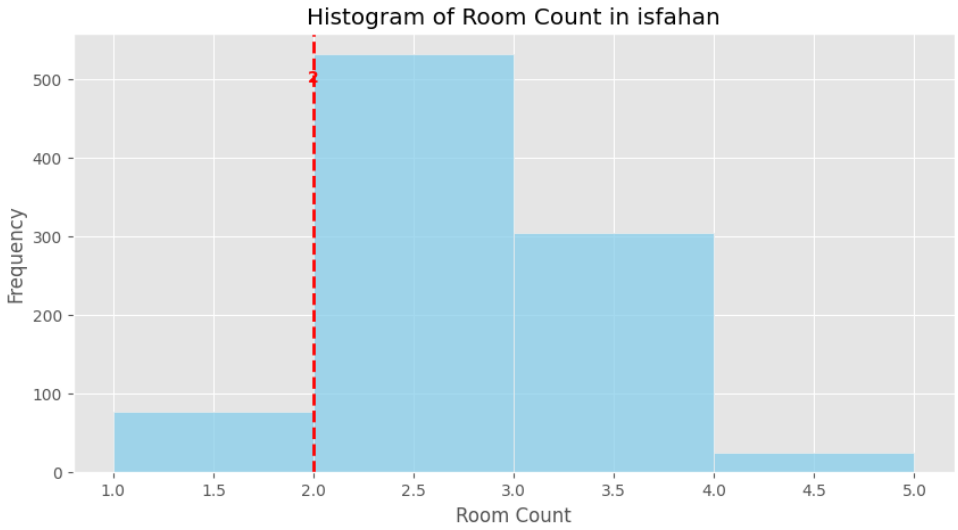
نمودار 21

که همانطور که از نمودار ها مشخص است، خانه های تازه ساخت و نوساز قیمت متراژی بالاتری نسبت به خانه های چند سال ساخت دارند. نکته ای که برای من جالب بود این بود که در شهر های تهران و اصفهان میانه خانه های آگهی شده کمتر از بقیه شهر ها بوده است و به نظرم دلیل این موضوع می تواند آن باشد که این دو شهر، پر جمعیت ترین کلانشهر های ایران هستند و احتمالا به همین دلیل و به دلیل تراکم شهر و ازدیاد واحد های آپارتمانی و همچنین به دلیل وجود بافت قدیمی تر، خانه های قدیمی تر هم در آگهی ها وجود زیادی دارند و به همین دلیل میانه سال ساخت خانه ها برای این دو شهر نسبت به بقیه کمتر است.

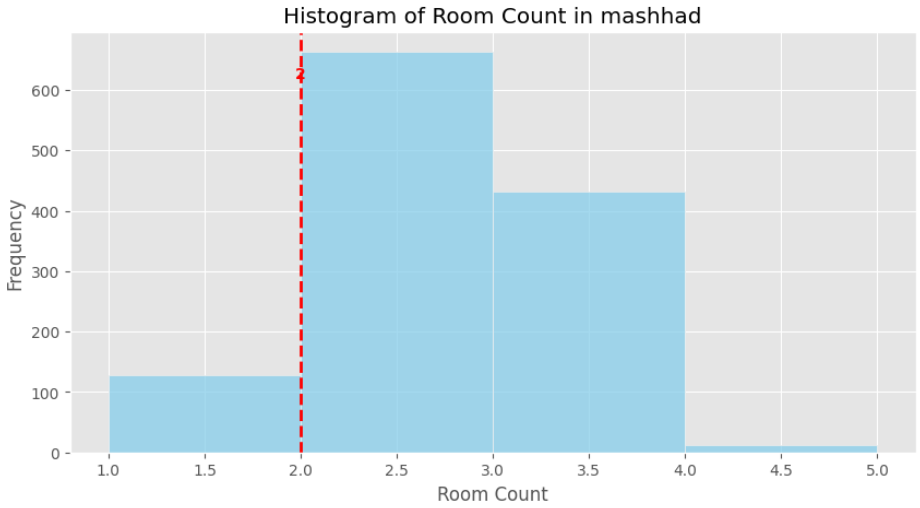
در مرحله بعد نیز مشابه مراحل قبلی توزیع تعداد اتاق ها را بررسی کردم:



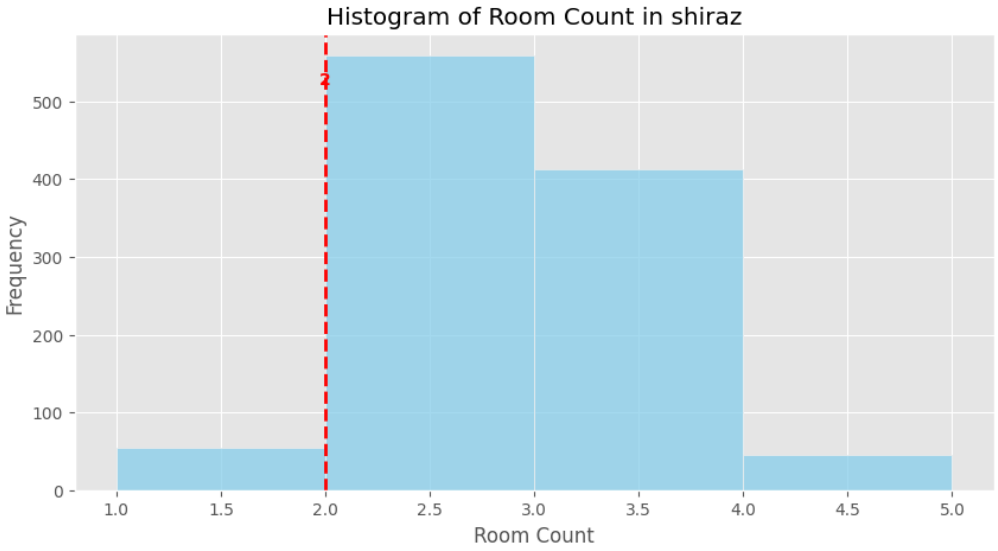
نمودار 22



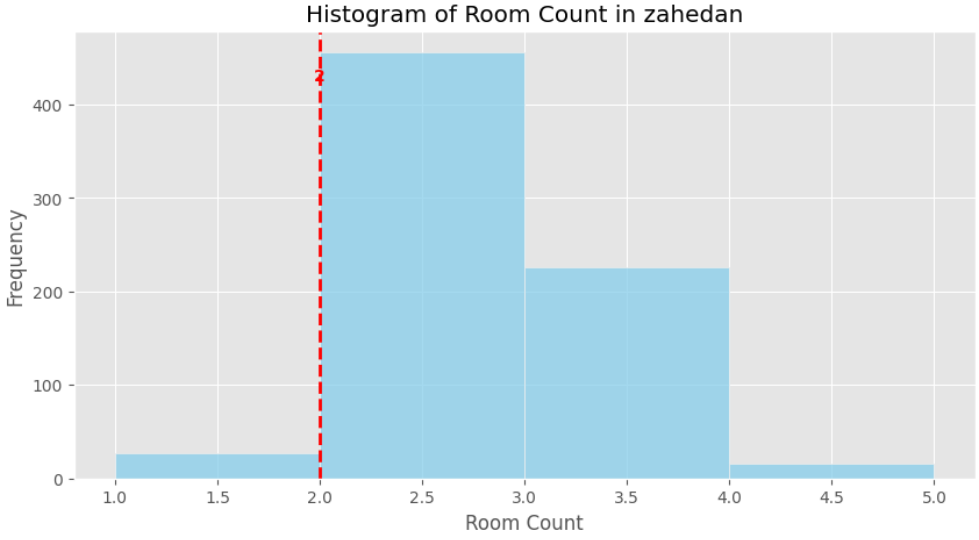
نمودار 23



نمودار 24



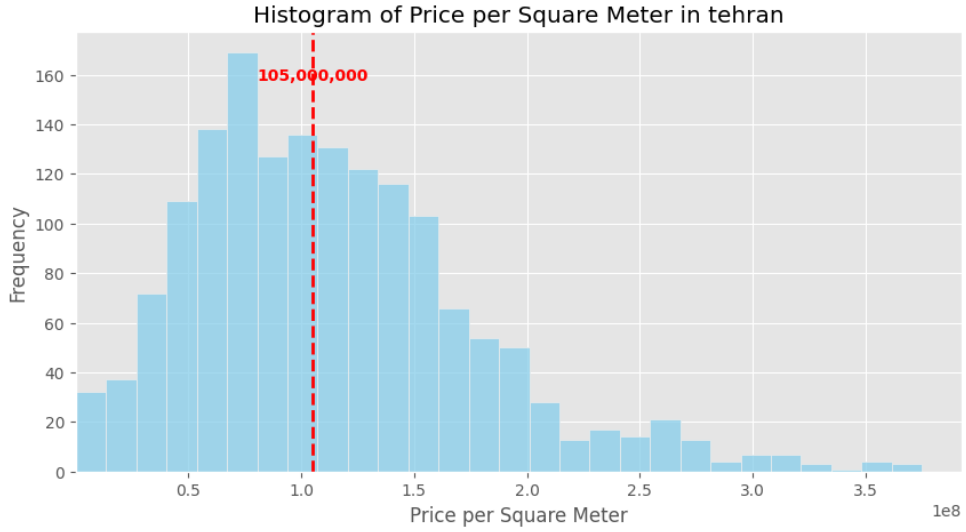
نمودار 25



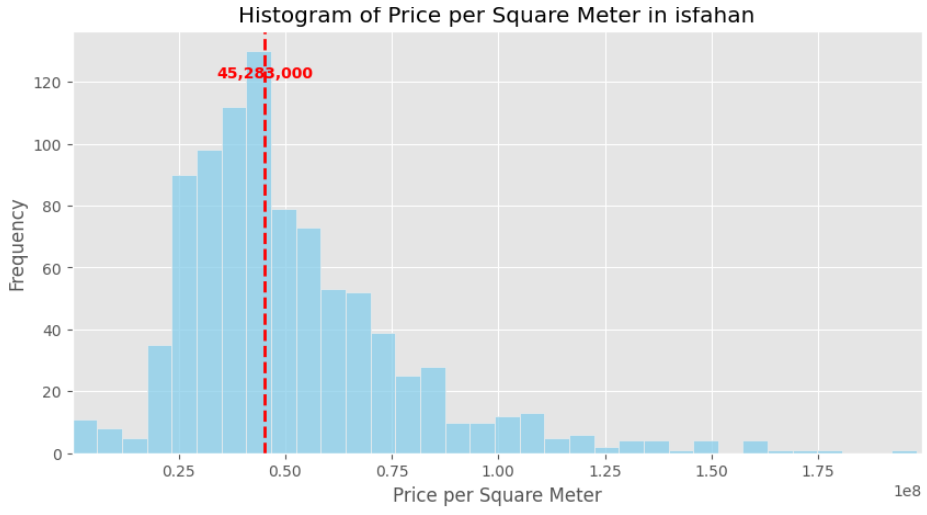
نمودار 26

همانطور که در نمودار ها مشخص است در همه شهر های مورد بررسی، بیشترین تعداد تکرار برای خانه های شامل دو اتاق هستند و این با واقعیت نیز همخوانی دارد چرا که به طور معمول در اکثر کلانشهر های تهران خانه های دو خوابه بیشترین تعداد را دارند.

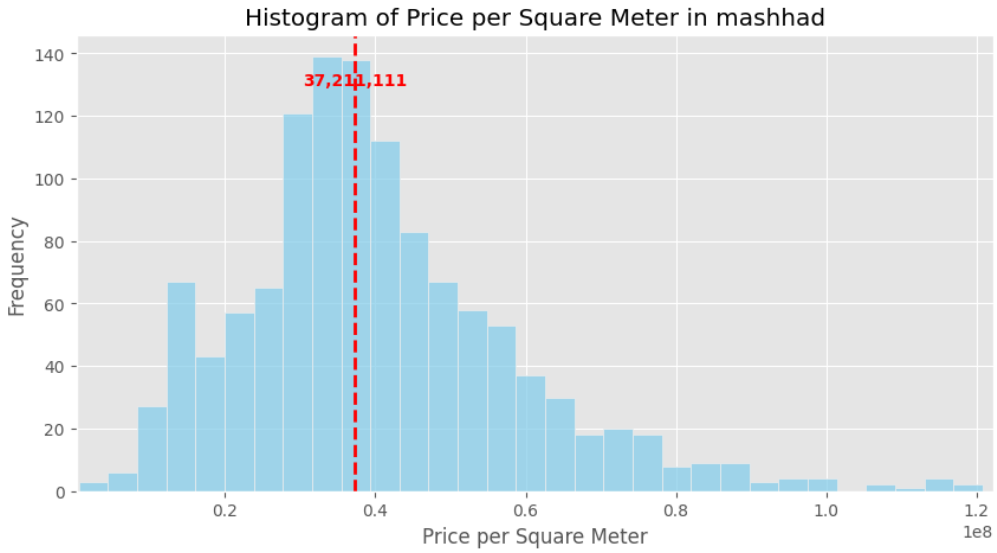
در ادامه نمودارهای قیمت بر حسب متر مربع برای هر شهر را رسم کردم، در این قسمت به دلیل بالا بودن scale و مقیاس قیمت ها اول سعی کردم با استفاده از نمودار لگاریتمی و یا نرمال سازی داده ها نمودار هر بخش رو رسم کنم اما نتیجه نداشت و شکل نمودار شکل درستی نبود، از طرفی با توجه به اینکه کمینه و بیشنیه قیمت ها در مقیاس لگاریتمی خیلی فاصله ای نداشتند، نمودار لگاریتمی هم نتیجه بخش نبود. به همین دلیل داده های پرت را بر اساس صدک ۱ و ۹۹ دور ریختم و مقیاس بندی نمودار را هم با استفاده از histogram_bin_edge انجام دادم که نتیجه به شرح زیر است:



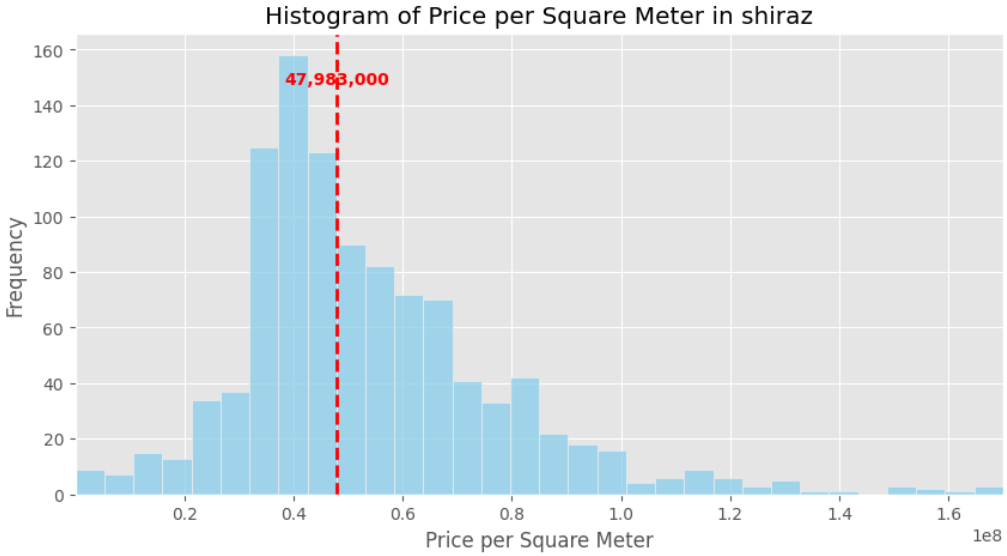
نمودار 27



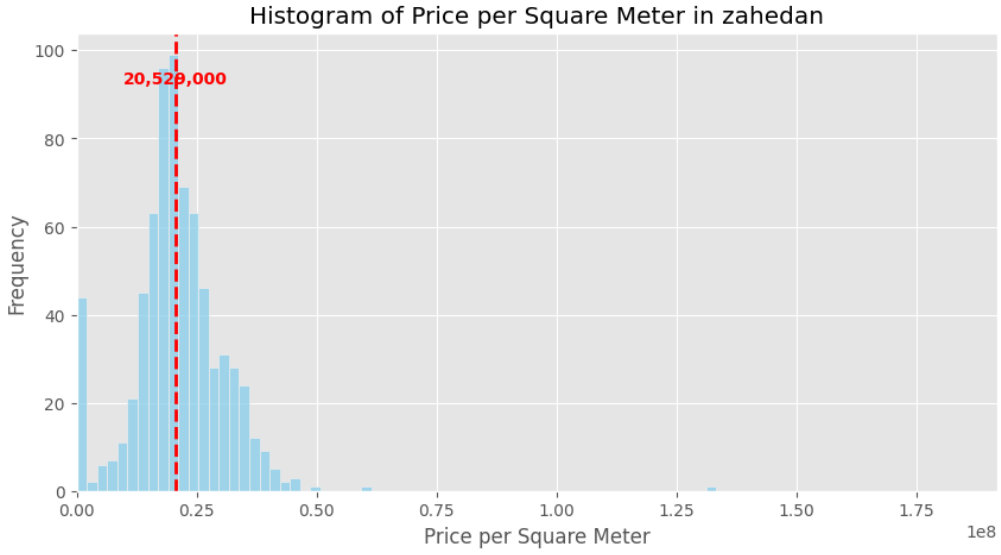
نمودار 28



نمودار 29



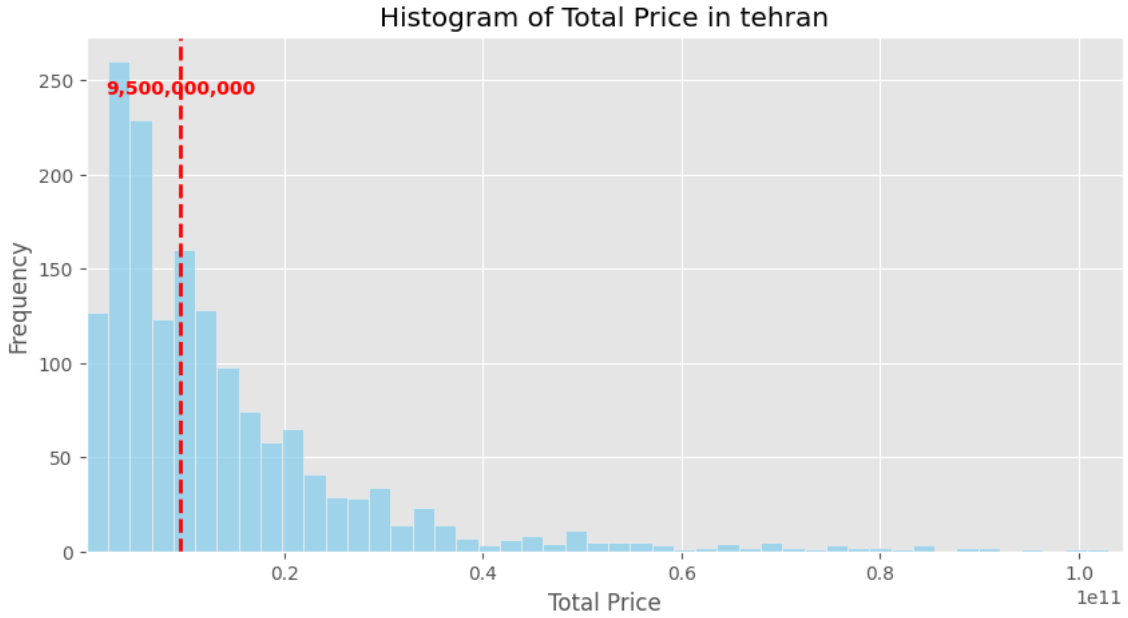
نمودار 30



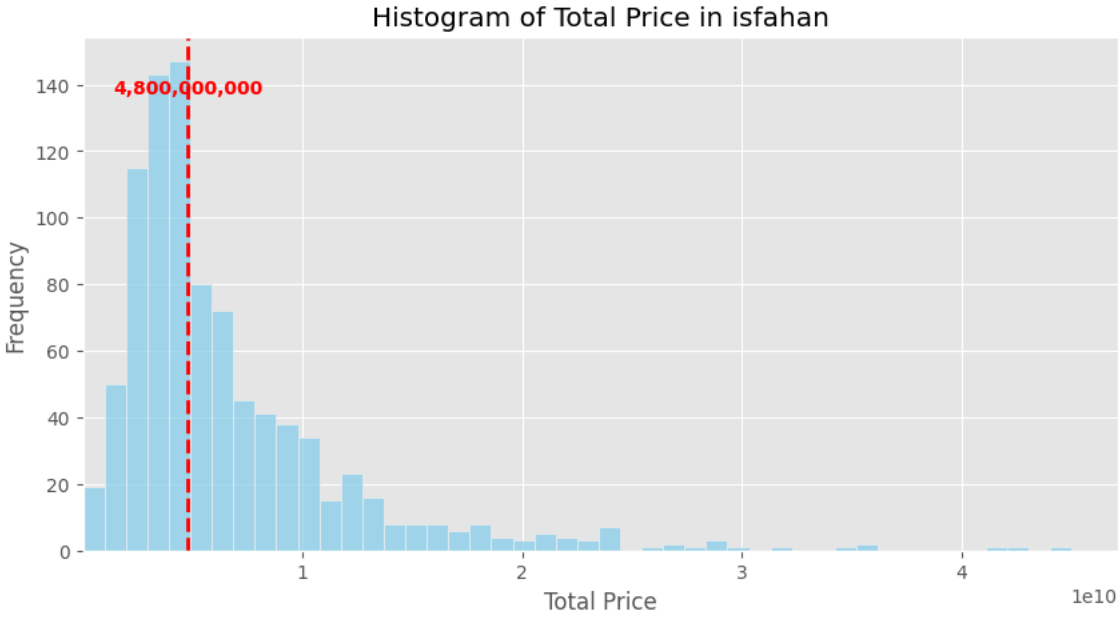
نمودار 31

که همانطور که قابل انتظار بود قیمت بر حسب متر خانه در شهر تهران بیشتر از بقیه شهر ها می باشد که با مشاهدات ما منطبق است.

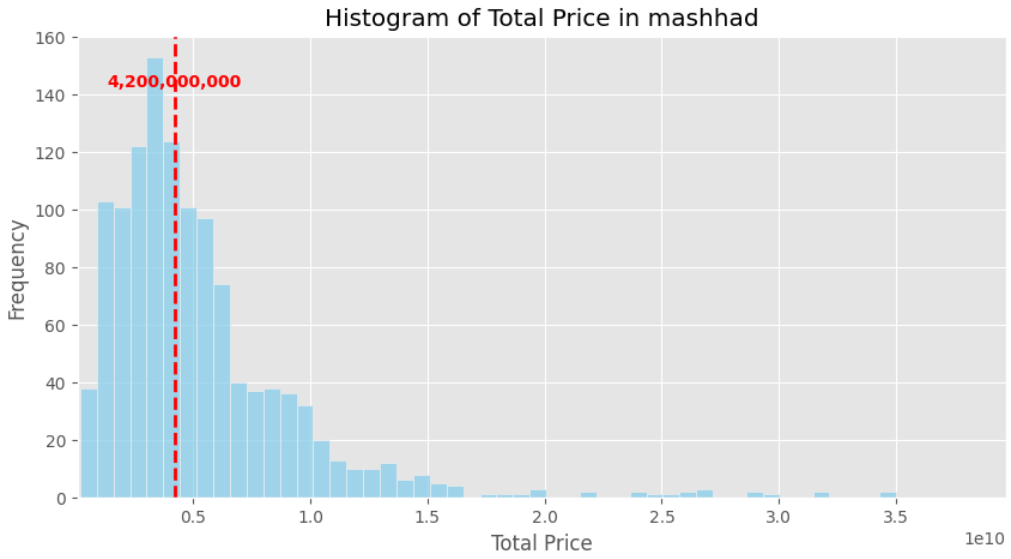
در مرحله بعد برای توزیع قیمت کلی در هر شهر هم مانند قسمت بالا عمل کردم و نمودار های حاصله به شرح زیر است:



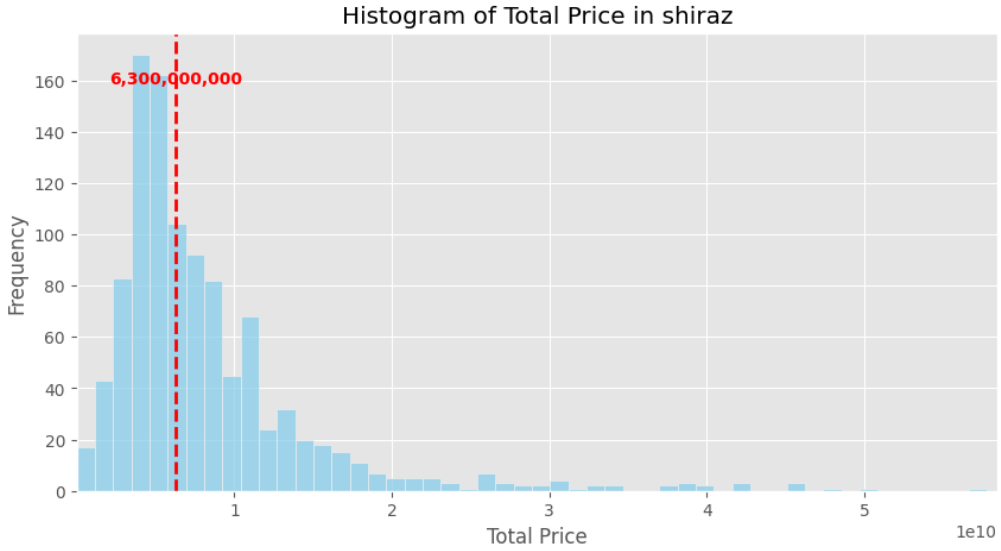
نمودار 32



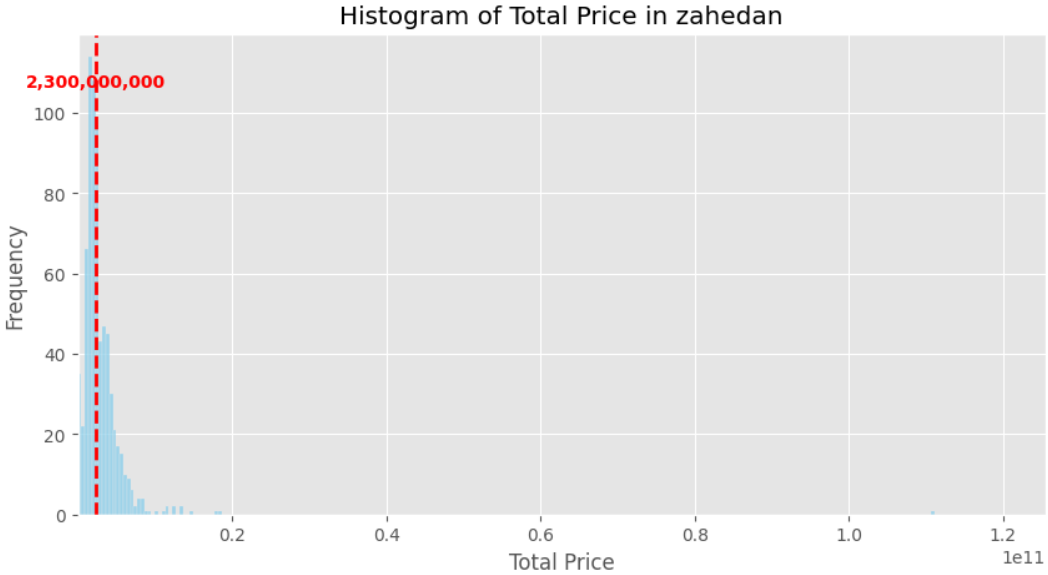
نمودار 33



نمودار 34



نمودار 35



نمودار 36

در این نمودار ها نیز همانطور که قابل انتظار است تهران بیشترین میانه قیمت کلی و زاهدان کمترین میانه قیمت کلی را دارد.

اظهارنامه استفاده از هوش مصنوعی

تأیید می‌کنم که از ابزارهای هوش مصنوعی مطابق با دستورالعمل‌های بارگذاری شده در سامانه Elearn درس به طور مسئولانه استفاده کرده‌ام. تمام اجزای کار خود را درک می‌کنم و آماده بحث شفاهی درباره آنها هستم.