نام و نام خانوادگی: درسا الهوئی

شماره دانشجویی: ۹۷۲۲۲۰۰۵

شماره تمرین: ۱,۲ **تاریخ تحویل: ۲۰ فروردین ۱۴۰۱**

عنوان تمرین: اعمال موارد مذکور در تمرین بر مجموعه داده مربوط به ساختمان های اجاره ای آلمان

---

۱. مقدمه

در این پروژه هدف انجام بررسی آماری داده های مربوط به ساختمان های اجاره ای کشور آلمان و ارائه مدلی جهت پیش بینی قیمت آنها می باشد.

۲. بررسی و پاکسازی داده ها

در گام اول بررسی کردیم که آیا اطلاعات جمع آوری شده در چند تاریخ مذکور، تکراری می باشند یا خیر. که نتیجه بررسی منفی می باشد و دادها تکراری نیستند. در گام بعد ستون هایی که در روند پیش بینی، مفید به نظر نمی رسیدند را حذف می کنیم که با این کار، تعداد ستون ها از ۴۹ به ۳۵ کاهش می یابد.

در ادامه با توجه به جدول توضیح داده (data.describe())، داده هایی که به نظر منطقی نمی رسیدند را حذف کردیم:

- ساختمانی که فضای آن ۰ مترمربع باشد به معنی است. بنابراین سطرهایی از داده که مقدار livingSpace در آنها برابر ۰ بود را حذف کردیم.
- جمع آوری داده ها در سال ۲۰۱۹ صورت گرفته است. بنابراین ساختمانهایی که سال ساخت آنها بعد از ۲۰۱۹ باشد، منطقی نیستند و در نتیجه آنها را حذف می کنیم.
- خانه ای که قیمت اجاره آن برابر ۰ باشد نیز بی معنی است. در نتیجه ساختمان های با این ویژگی را حذف می کنیم.

پس از این گام، به بررسی درصد داده های null در هر ستون می پردازیم.

با توجه به آنچه در توضیح ستون ها گفته شد، totalRent از جمع baseRent، serviceCharge و heatingCost بدست می آید. بنابراین با به کار گیری این نکته درصدی از مقادیر null ستون totalRent را پر می کنیم و باقی آنهایی که مقدار totalRent در آنها برابر ۰ است را حذف می کنیم چرا که این ستون، هدف ما برای پیش بینی است.

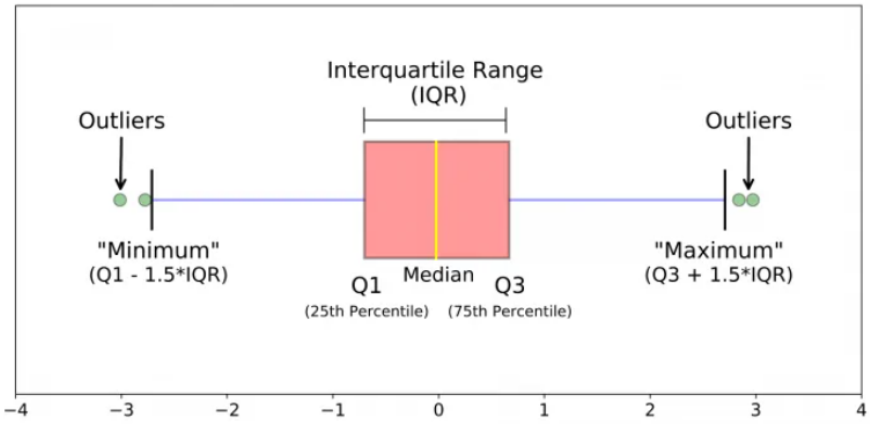
ستون هایی که در آنها مقدار null بیش از ۳۵٪ می باشد را حذف می کنیم.

regio1	0.000000
serviceCharge	1.693259
heatingType	16.353496
telekomTvOffer	11.461579
newlyConst	0.000000
balcony	0.000000
telekomUploadSpeed	11.728693
totalRent	0.000000
yearConstructed	21.288605
hasKitchen	0.000000
cellar	0.000000
baseRent	0.000000
livingSpace	0.000000
condition	25.439564
lift	0.000000
baseRentRange	0.000000
typeOfFlat	13.452563
noRooms	0.000000
floor	17.504141
noRoomsRange	0.000000
garden	0.000000
livingSpaceRange	0.000000
regio2	0.000000
regio3	0.000000

ستون condition را که ۲۵٪ آن مقدار null دارد با مقدار other پر می کنیم. چند ستون telekomTvOffer، heatingType، typeOfFlat دیگر را که در کد ذکر شده است، به این روش با مقدارهای unk یا other پر می کنیم. مقادیر null ستون های serviceCharge، telekomUploadSpeed، yearConstructed و floor را با مقدار میانگین ستون مربوطه پر می کنیم.

در گام بعد به حذف داده های پرت می پردازیم i; این کار را با به کارگیری روش IQR صورت می گیرد. با مشخص کردن چارک ها و محاسبه IQR (IQR = Q3 – Q1)، بازه ی نرمال داده ها را از Q1 – 1.5*IQR تا Q3 + 1.5*IQR در نظر می گیریم و به این طریق سایر داده های خارج از این بازه حذف می شوند. در نتیجه این گام، تعداد داده ها از ۱۶۰۰ به ۱۲۰۸ داده کاهش می یابد.

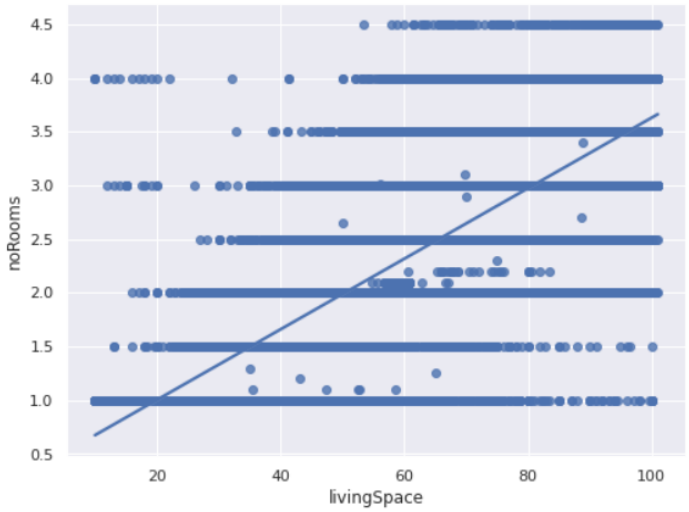
سپس به روش IQR داده های پرت را نیز حذف می کنیم.



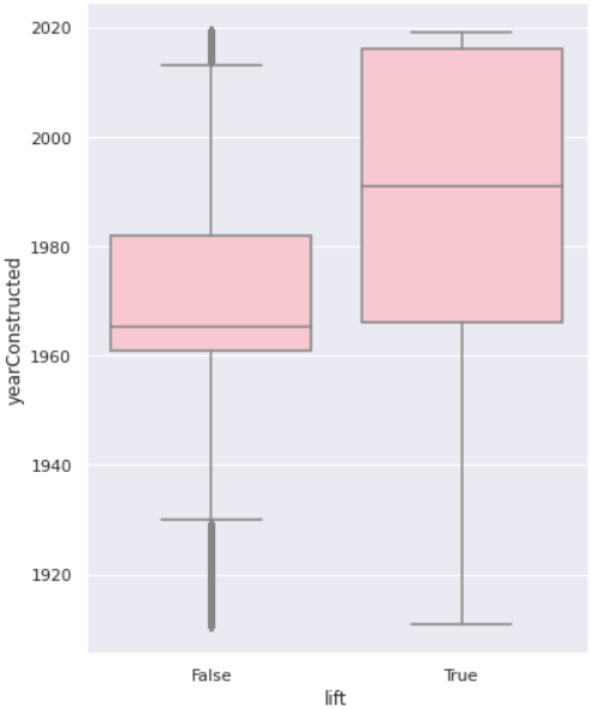
قصد داشتم با مدل Bert از ستون description نیز استفاده کنم اما فصت نشد.

۳. تصویرسازی و بررسی آماری داده ها

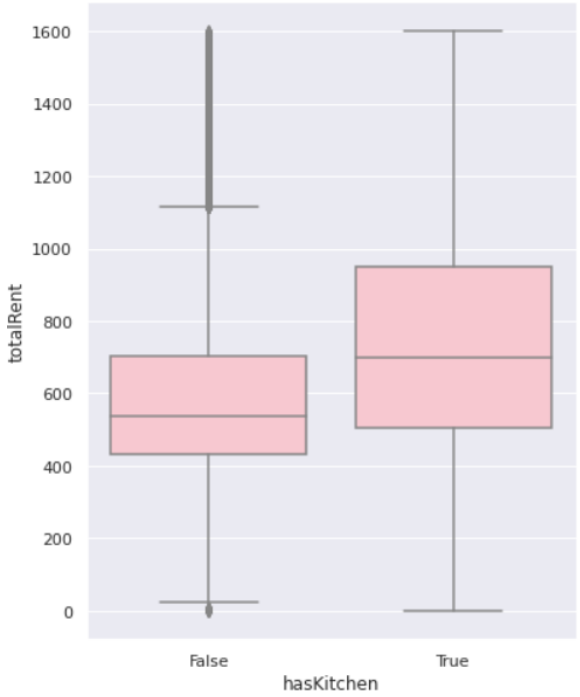
نمودار زیر رابطه بین تعداد اتاق ها و فضای ساختمان را نشان می دهد. می توان دریافت که تعداد اتاق های بیشتر در خانه با متراژ بیشتر یافت می شوند.



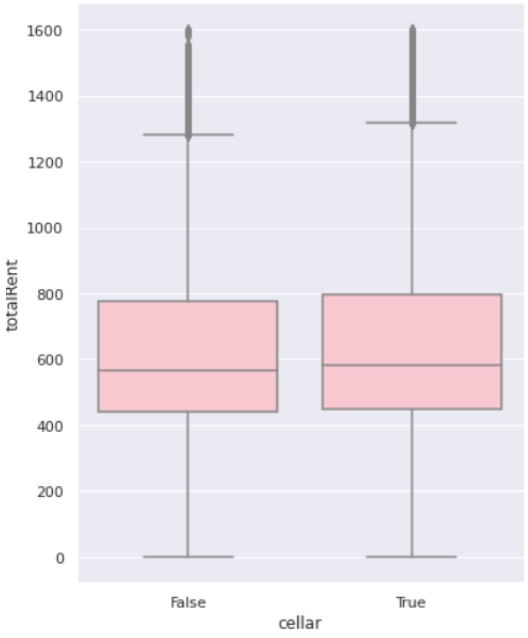
با توجه به نمودار جعبه ای زیر می توان دریافت که تعداد بیشتری از خانه هایی که جدیدتر ساخته شده اند نسبت به ساختمان های قدیمی تر از آسانسور بهره می برند.

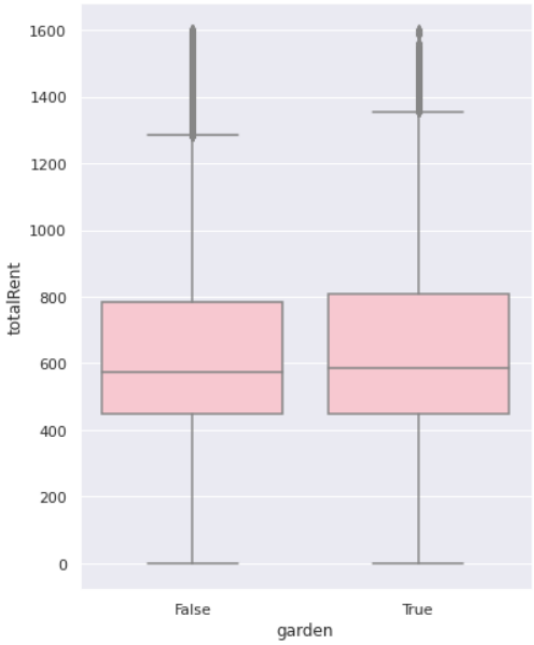


همچنین خانه هایی که آشپزخانه آماده دارند، رنج قیمت بالاتری نیز دارند.

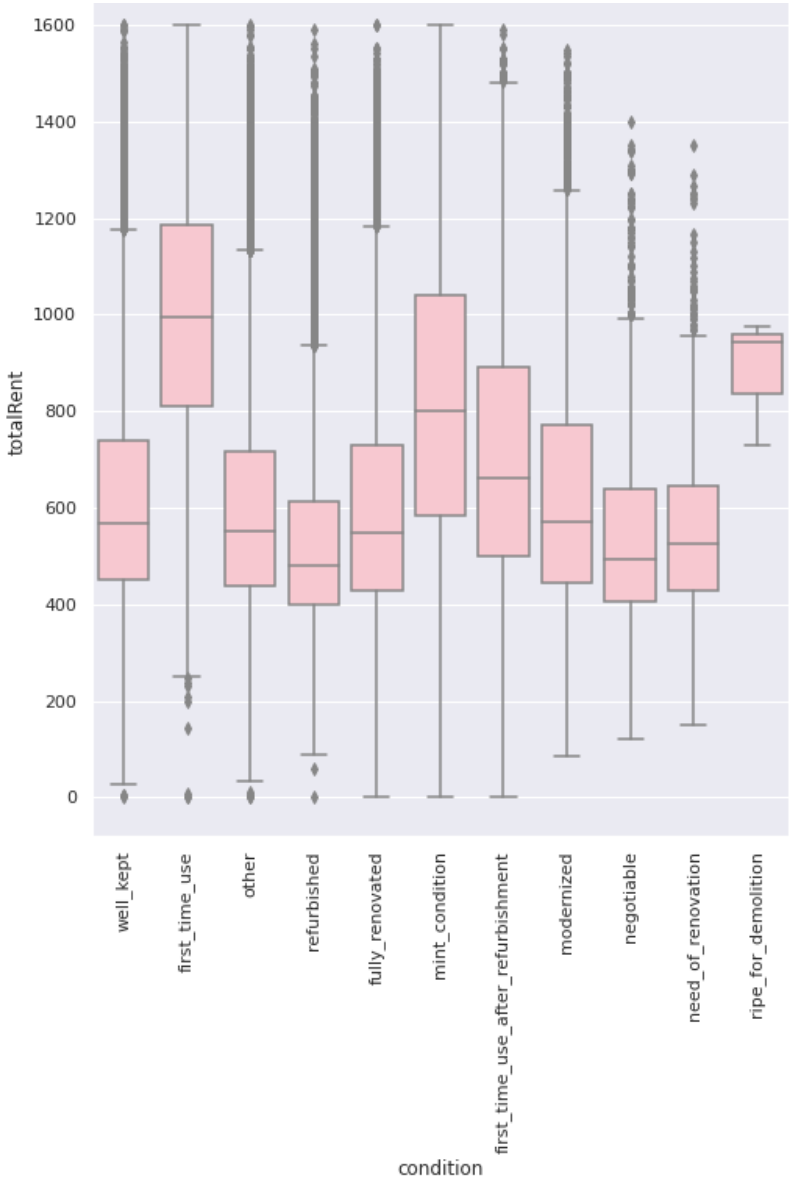


طبق دو نمودار زیر داشتن باغچه و انبار تاثیر بسزایی در قیمت ندارد.

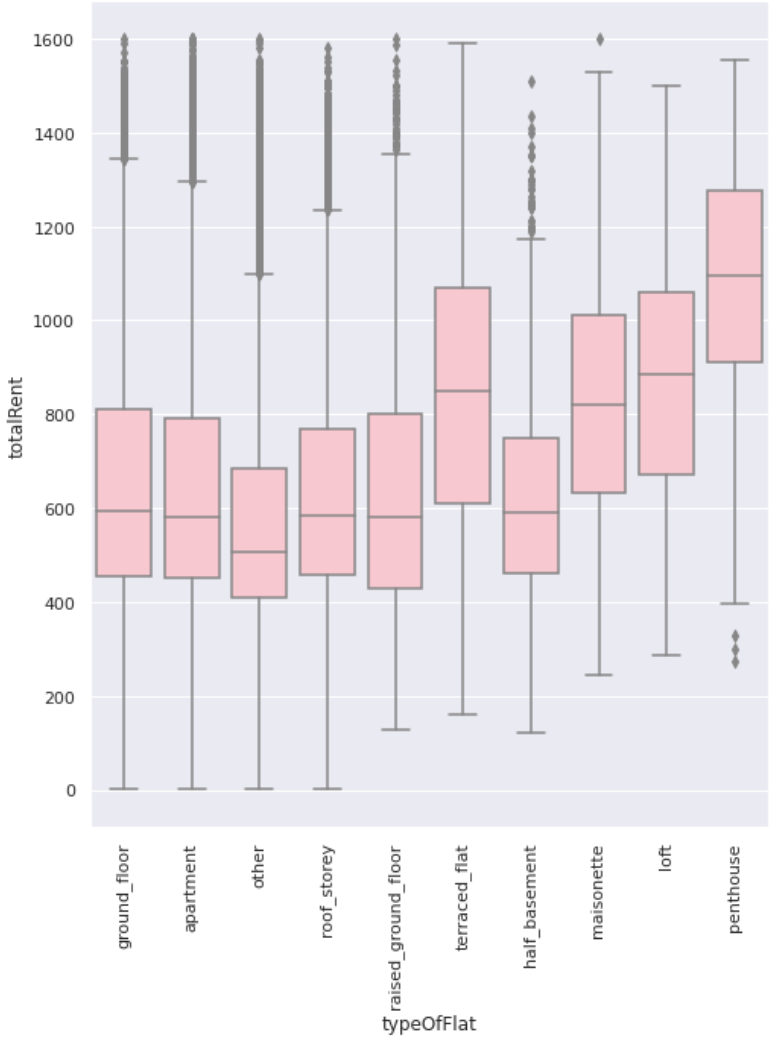




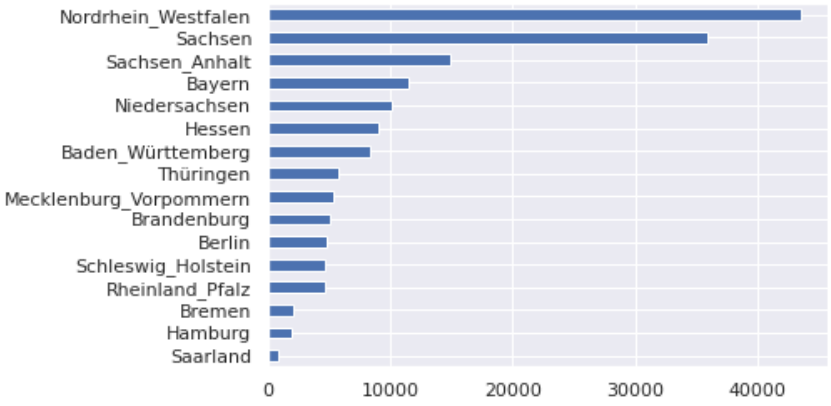
هرچه وضعیت ساختما بهتر باشد، رنج قیمتی آن نیز بالاتر است. همانطور که مشاهده می شود ساختمانی که برای اولین با استفاده می شود به وضوح رنج قیمتی بالاتری دارد.



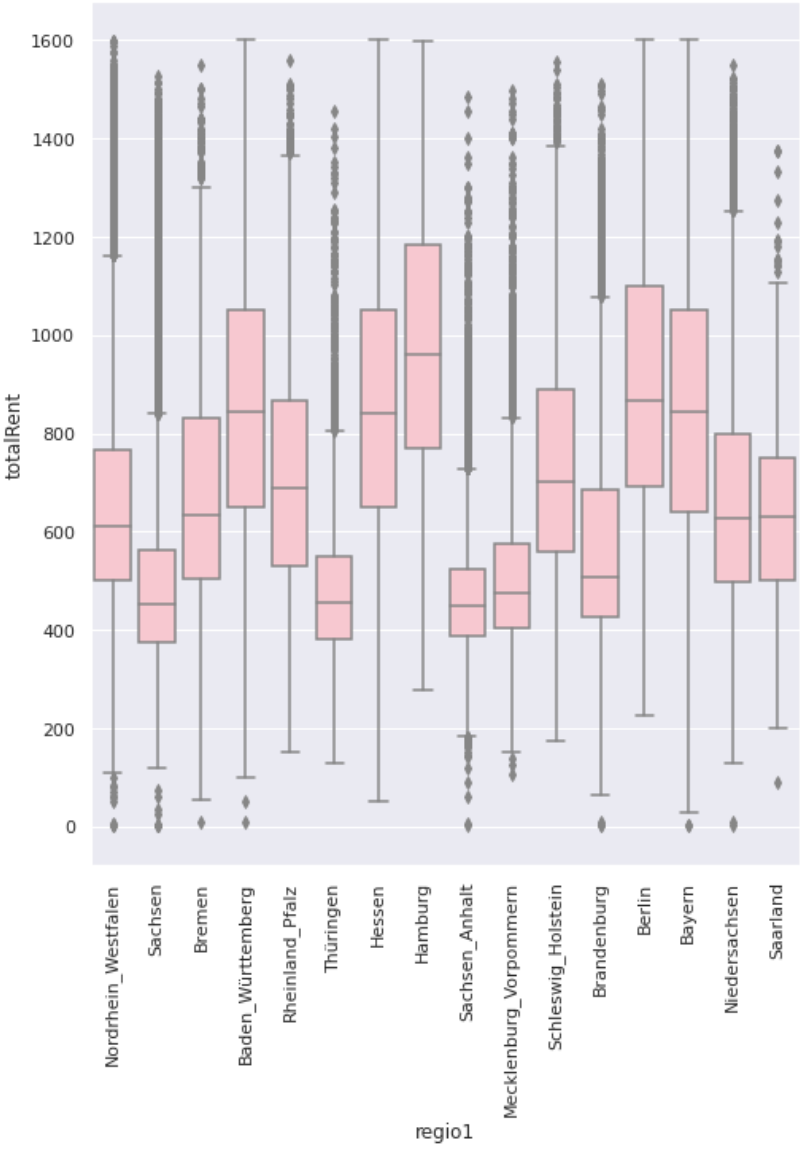
همچنین نوع ساختمان نیز در قیمت آن نقش دارد. مشاهده می شود که رنج قیمت پنت هاوس به نسبت از بقیه بالاتر است.



نمودار زیر نیز تعداد آگهی های ساختمان را در بخش های مختلف کشور نمایش می دهد.



همچنین با استفاده از نمودار جعبه ای، پرکندگی قیمت در هر بخش نیز مشاهد میشود.



پس از این مرحله برای تبدیل داده های کیفی یا categorical به داده کمی از get_dummies استفاده می کنیم و آنها را مقیاس بندی کرده و با اعمال الگوریتم PCA آنها را برای دادن به مدل آماده می کنیم.