نام ونام خانوادگی : سارا رضایی
شماره دانشجویی: ۹۸۲۲۲۰۴۳
عنوان : گزارش و تحلیل دیتاست دوم تمرین ۱ درس مبانی یادگیری ماشین

مقدمه

داده های مورد بررسی در این تمرین داده های یکی از بزرگترین آژانس های املاک آلمان است. این دیتاست فقط شامل خانه های اجاره ای است.

مجموعه داده شامل بسیاری از ویژگی های مهم، مانند اندازه منطقه زندگی، اجاره، هر دو اجاره پایه و همچنین اجاره کل ، مکان (خیابان و شماره خانه ، کد پستی ، ایالت ، منطقه و ...)، نوع انرژی وفیچر های دیگر می باشد.

در این دیتاست بررسی خواهیم کرد که کدام مناطق بیشترین قیمت خانه را دارند. کدام ویژگی ها تاثیربیشتری روی قیمت خانه دارد. همچنین در نهایت تلاش برای مدلسازی برای پیش بینی قیمت اجاره انجام خواهد شد.

متد ها و مدل ها

از بلاک اول تا قبل از هدینگ تسک ۱ لود و بارگذاری داده ها از کگل انجام شده و shape و head دیتافریم برای مشاهده ی کلی داده ها چاپ شد. همچنین در بلاک آخر دیتافریم در یک دیتافریم دیگر کپی شده تا دیتافریم بدون تغییری برای اعمال پاک سازی با daskموجود باشد.

تسک اول preprocessing داده هاست که از **بلاک ۱۲** انجام شده است. در **بلاک ۱۲** درصد داده های null در هر ستون محاسبه شده است و به ترتیب صعودی نمایش داده شده است. همچنین زمان اجرای این سلول نیز محاسبه شده تا بعدا با کارایی در حالت استفاده از dask مقایسه شود.

در **بلاک۱۳و۱۴** بعدی ستون هایی که درصد داده های null در آنها بیشتر از ۳۰درصد است از دیتافریم حذف شده اند.

در **بلاک ۱۶** سطر هایی که اطلاعات خانه ها بدون اجازه نهایی ثبت شده اند از دیتافرم حذف شده اند زیرا در مدلسازی نمیتوان از آنها استفاده کرد.

در **بلاک ۱۷** ستون هایی که شامل داده هایی هستند که اطلاعات غیرمفید یا غیرقابل استفاده دارند(مثلا داده هایی که نیاز به متن کاوی دارند) حدف شده اند.

در **بلاک ۱۸ و ۱۹** سطر های تکراری(dublicated) شناسایی و حذف شده اند.

در **بلاک ۲۰** داده های عددی ای که همچنان null مانده اند با میانگین داده های ستون خودشان پر شدند.

در **بلاک ۲۱** داده های categorical ای که همچنان null مانده اند با پرتکرار ترین داده ی ستون خودشان پر شدند.

در **بلاک** ۲۳ داده های پرت در ستون داده های عددی با تکنیک IQR شناسایی و حذف شدند.

تسک دوم Exploratory Data Analysis که **از بلاک ۲۴** انجام شده است. در این قسمت اطلاعات آماری فیچر ها با استفاده از دستور describe نمایش داده شده و تایپ داده های هر ستون نیز بررسی شده. همچنین ارتباط فیچر ها در نمودار های مختلف بررسی شده است که در بخش نتایج بیان میشود.

در این بخش تعداد آگهی های مربوط به هر منطقه در بلاک ۲۵ نمایش داده شده.

تسک سوم مدل سازی برای پیش بینی قیمت اجاره ی خانه بر اساس فیچر های موجود میباشد که از **بلاک ۳۹** آماده سازی داده ها برای دادن به مدل ها انجام شده است.

در **بلاک ۴۵** داده ها عددی scale شدند و در **بلاک ۴۶** داده های categorical با متد ordinal encoder برای دادن به مدل کدگذاری شدند.

در **بلاک ۴۸** داده های trainو test به نسبت ۲۰ ۸۰ جدا شدند.

اولین مدلی که آموزش داده شد linear regression میباشد که score آن برابر با ۰,۷۵ میباشد.و mean absolute error برابر ۱۲۲ میباشد.

دومین مدلی که آموزش داده شد polynomial regression میباشد که با درجه چند جمله ای برابر ۳ score آن برابر با ۰,۸۰ میباشد و mean absolute error برابر ۱۰۶ میباشد.

سومین مدلی که آموزش داده شد decision tree میباشد که score آن برابر با ۰,۷۰ میباشد و mean absolute error برابر ۱۲۲ میباشد.

decision tree با مقادیر `random_state=0 , max_depth= 10` نیز آموزش داده شد که score آن برابر با ۷۸. ۰ میباشد و mean absolute error برابر ۱۱۰ میباشد.

همچنین decision tree با مقادیر `random_state=50 , max_depth= 1000 , max_leaf_nodes=1000` نیز آموزش داده شد که score آن برابر با ۷۹۸. ۰ میباشد و mean absolute error برابر ۱۰۶ میباشد.

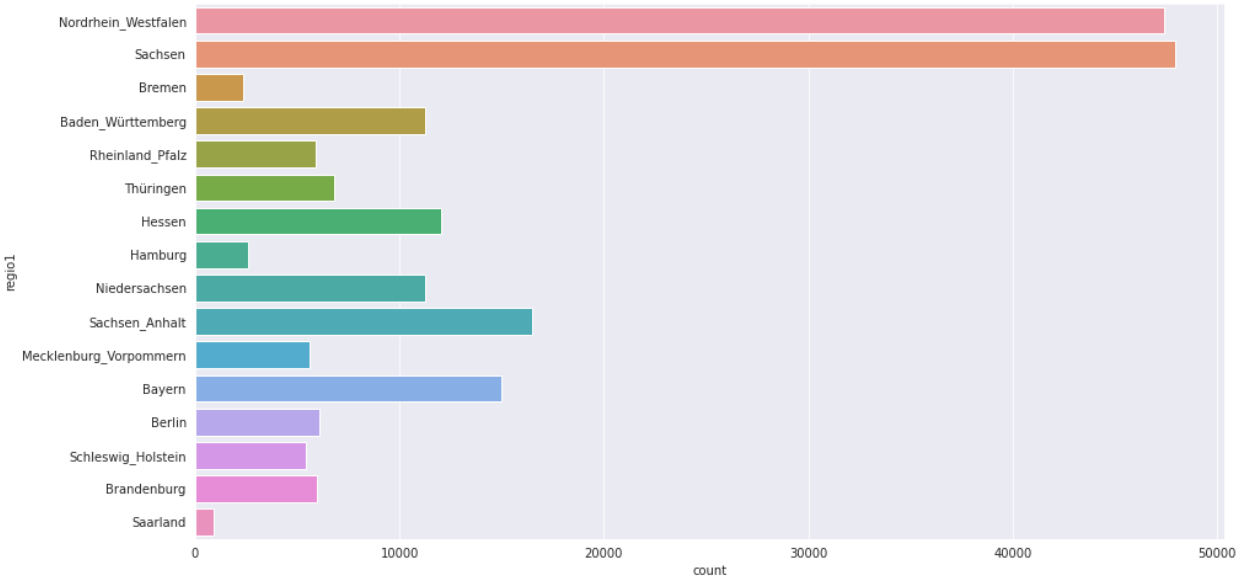
مدل چهارم XGBRegressor است که score آن برابر با ۸۲. ۰ میباشد و mean absolute error برابر ۹۹ میباشد. که بهترین نتیجه بین مدل ها میباشد.

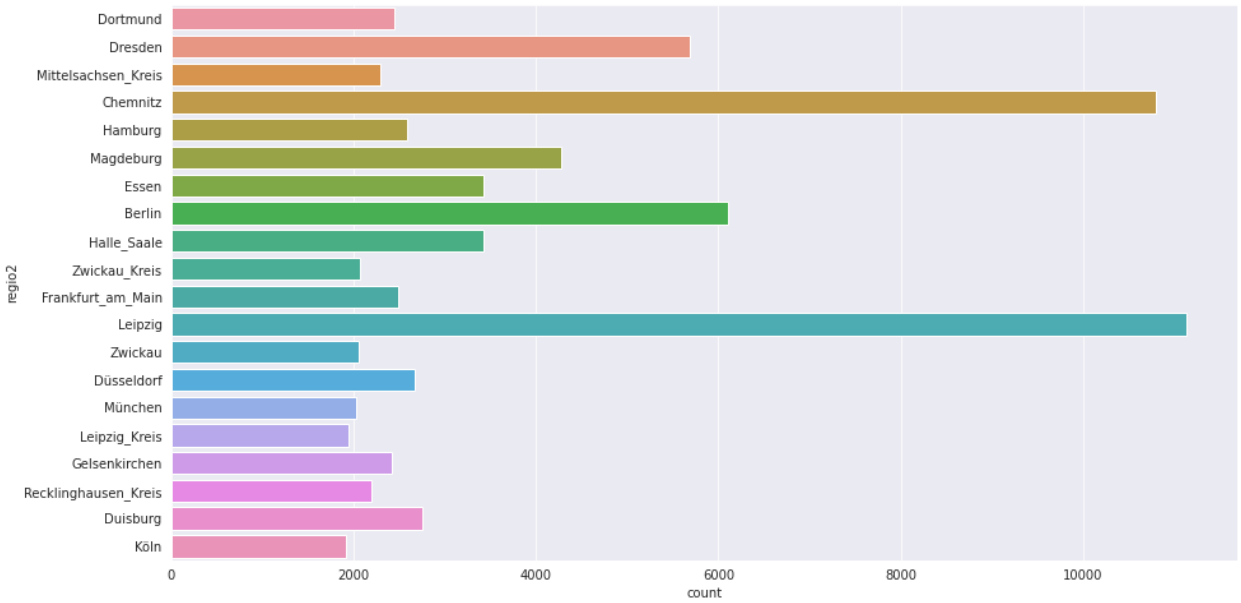
تسک چهارم استفاده از MULTIPROCESSING در پیش پردازش داده ها بود که برای شمارش داده های null و پر کردن داده های numerical با میانگین ستون متناظر از آن استفاده شد. و زمان های آنها نیز بررسی شد که نسبت به حالت عادی کمی افزایش یافته بود

تسک پنجم استفاده از dask در پیش پردازش داده ها بود که برای شمارش درصد داده های null از آن استفاده شد و مدت زمان اجر نسبت به حالت عادی کاهش یافته بود.

نتایج نمودار ها

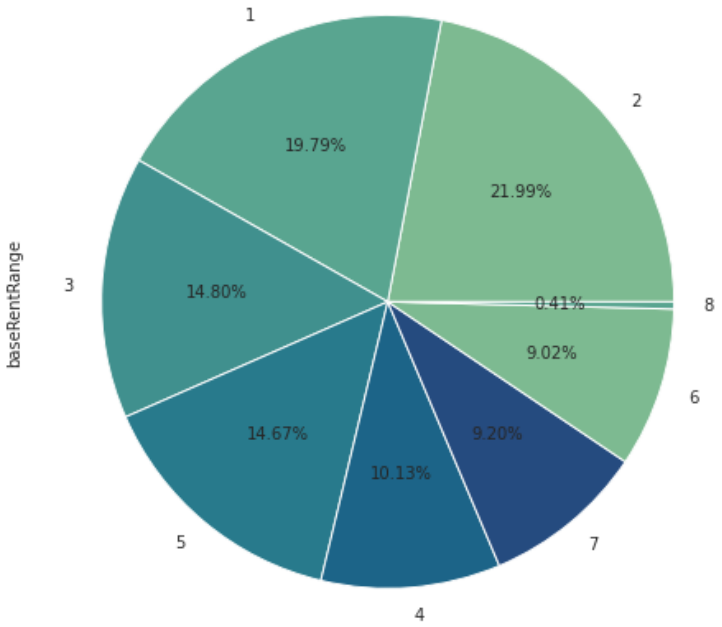
در دو نمودار زیر تعداد آگهی های در تعدادی از پرتکرار ترین ایالات و مناطق آلمان رسم شده است.

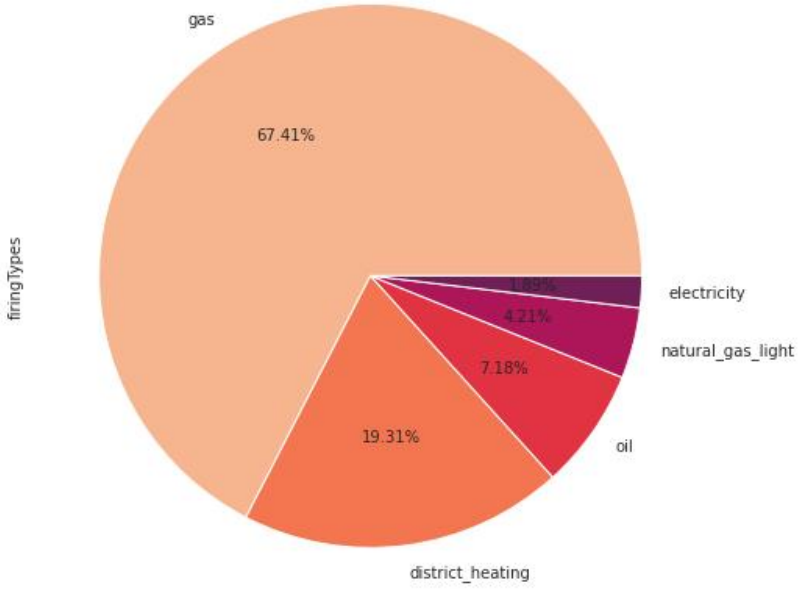


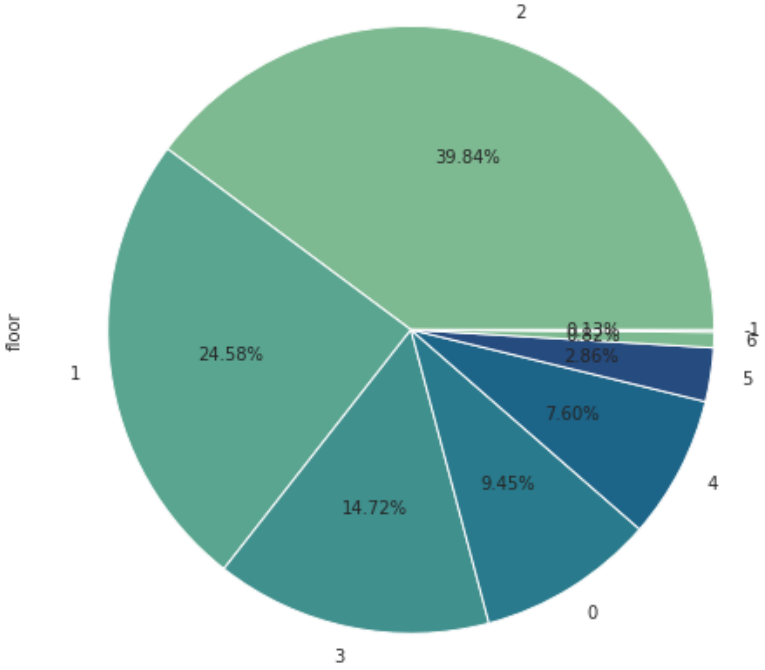


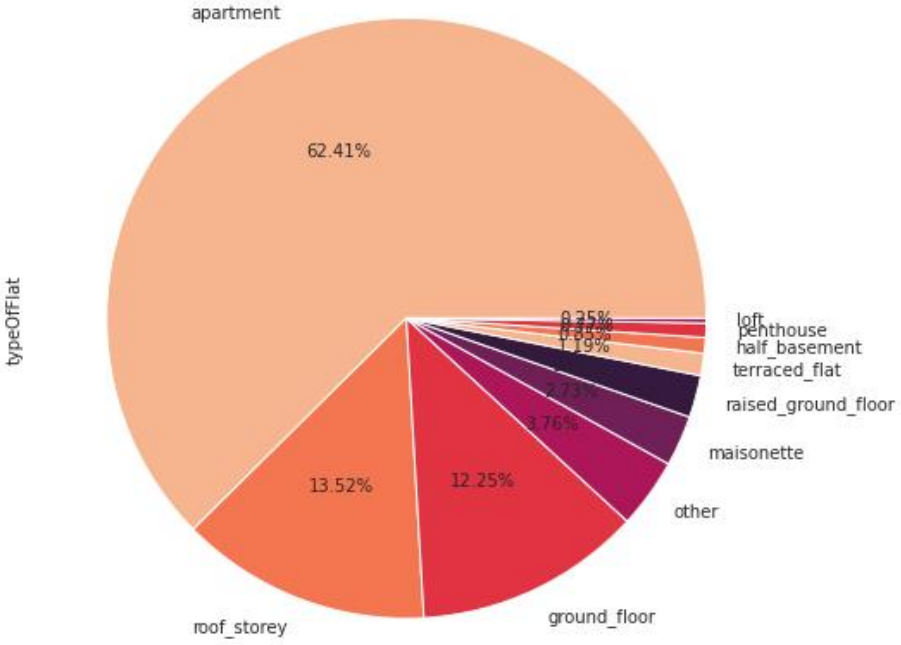
بر طبق این دو نمودار شهر Sachsen و ایالت Leipzig بیشترین تعداد آگهی را دارند.

نمودار هایی که در ادامه نمایش داده شده توزیع مقادیر مختلف داده ها ی categorical را نمایش میدهد

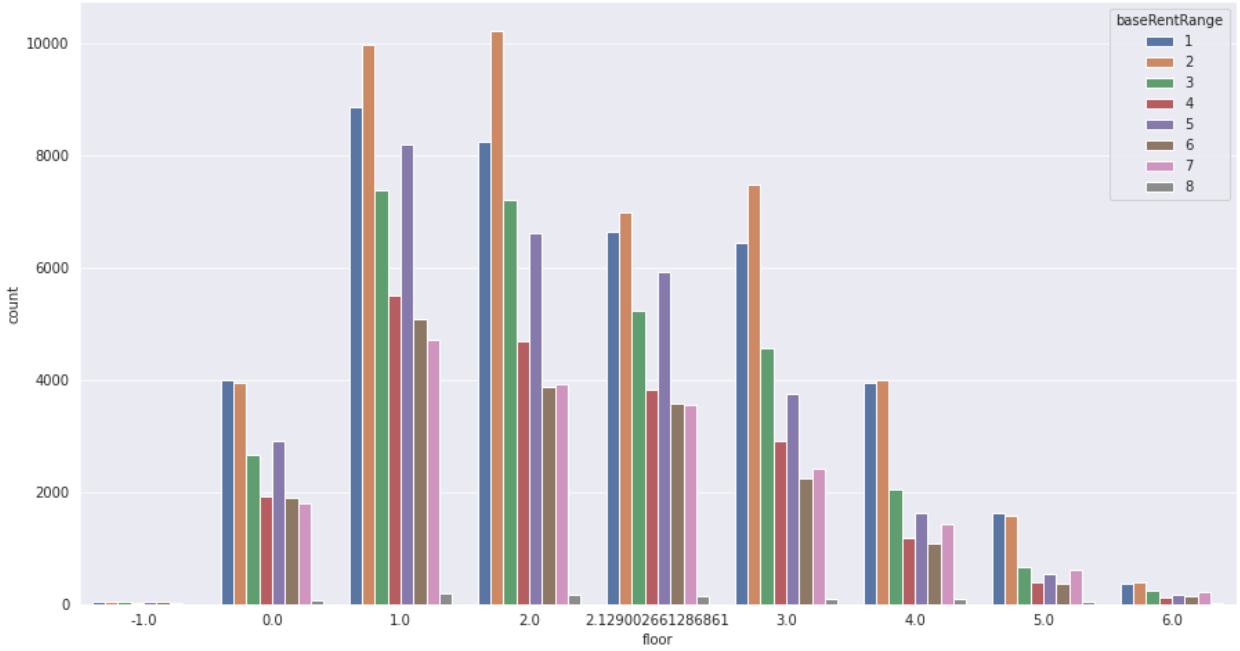




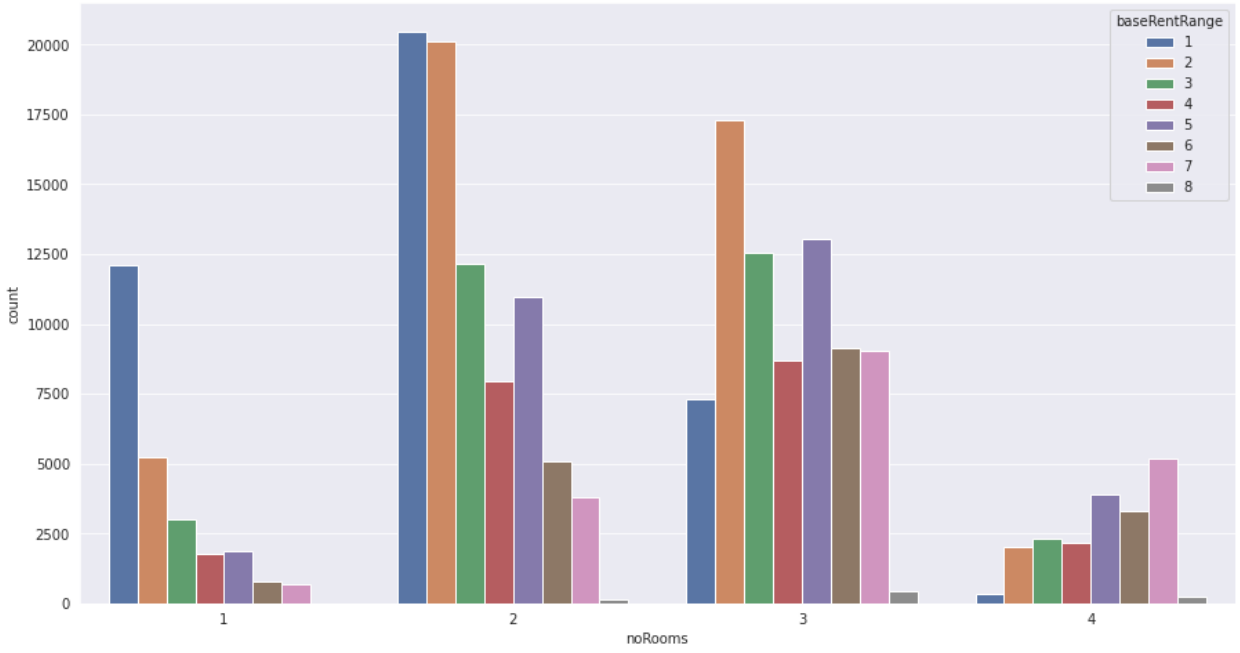

floor
2
39.84%
1
24.58%
3
14.72%
0
9.45%
4
7.60%
5
2.86%
6
7
1


typeOfFlat
apartment
62.41%
roof_storey
13.52%
ground_floor
12.25%
other
maisonette
raised_ground_floor
terraced_flat
half_basement
penthouse
loft

نمودار زیر رابطه تعداد طبقات و قیمت پایه را نمایش میدهد که بر طبق آن خانه های طبقه اول بیشترین تعداد آگهی در کلاس ۷ که از کلاس های گران تر است را دارا میباشند



نمودار زیر رابطه تعداد اتاق ها و قیمت پایه را نمایش میدهد که بر طبق آن خانه های دارای ۳ اتاق تعداد بیشتری از آگهی های کلاس های گران تر را دارا میباشند



و در نهایت نمودار correlation را داریم که بر طبق آن فضای خانه و تعداد اتاق ها و سال ساخت بیشترین تاثیر را روی قیمت دارند