غزل دانایی -97222034

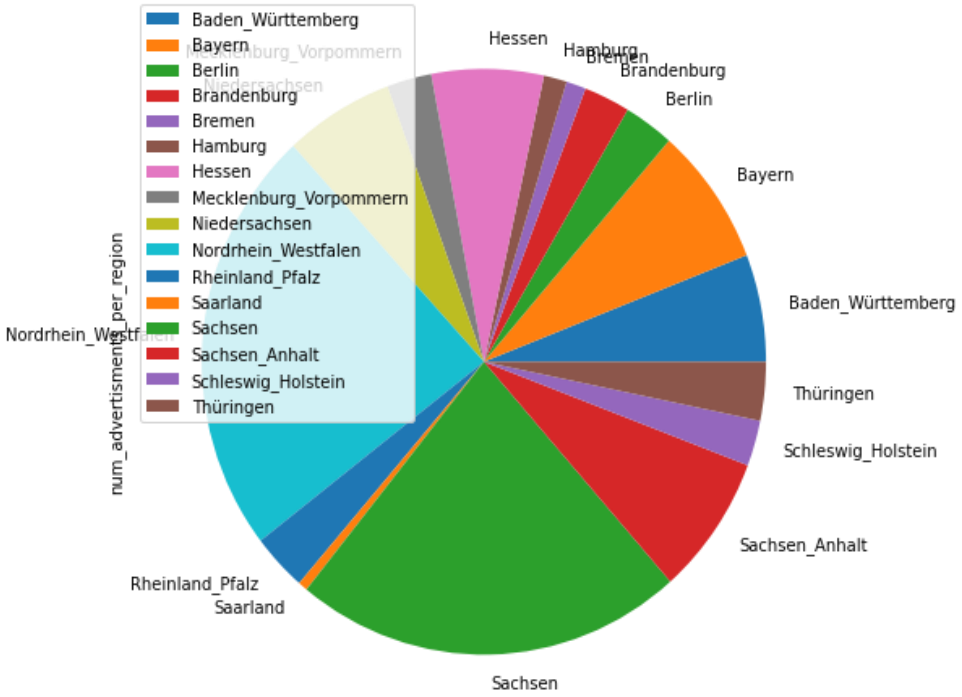
گزارش مجموعه داده دوم

سوال1-جزییات بیشتر در نوت بوک

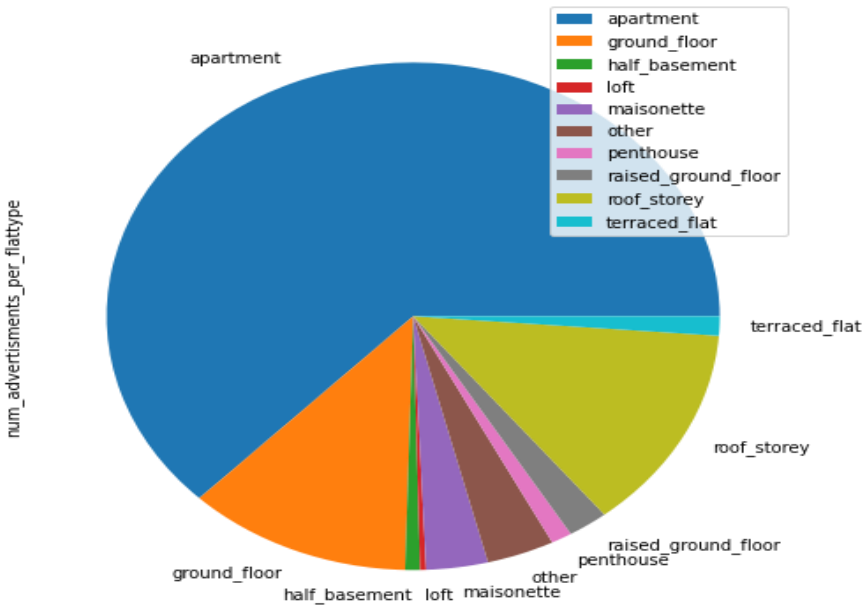
در ابتدا ستون های غیر مفید حذف شدند.ستون هایی که بیش از 50 درصد داده پوچ داشتند شناسایی و حذف شدند.داده های پوچ با مد یا میانگین جایگزین شدند و نیز داده ها پرت با متود سه برابر انحراف معیار شناسایی و حذف شدند.

سوال2-

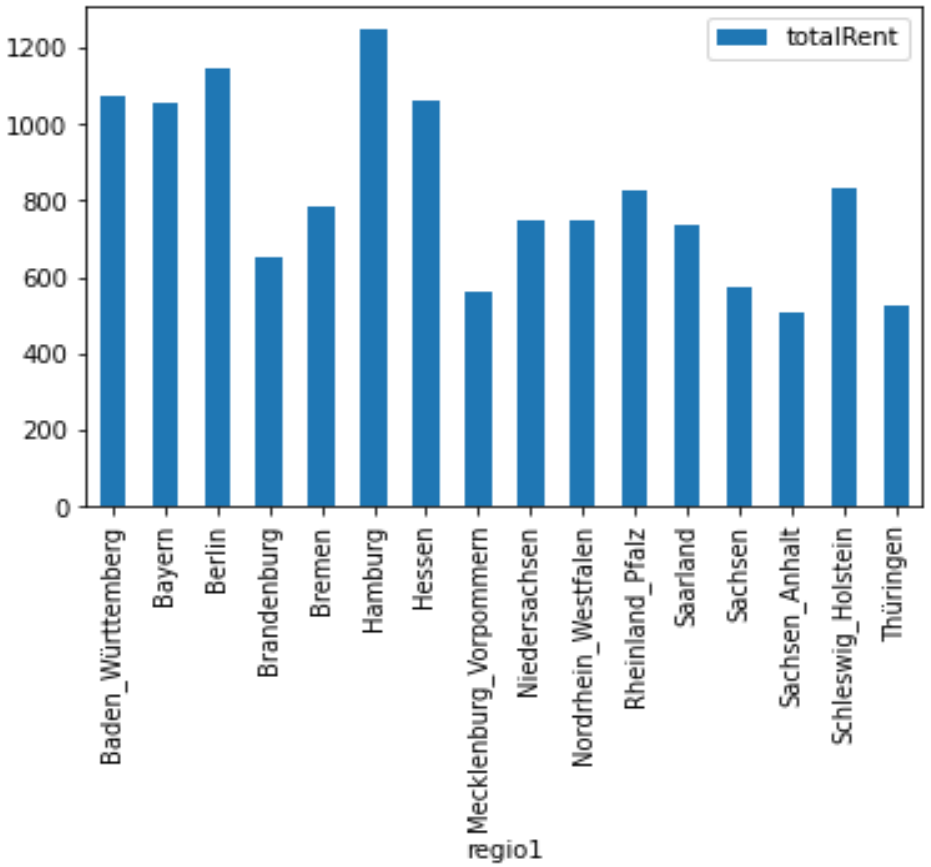
- در بین مناطق از نوع 1 آگهی شده بیشترین تعداد آگهی مربوط به Nordrhein_Westfalen و کمترین مربوط به Saarland می باشد.



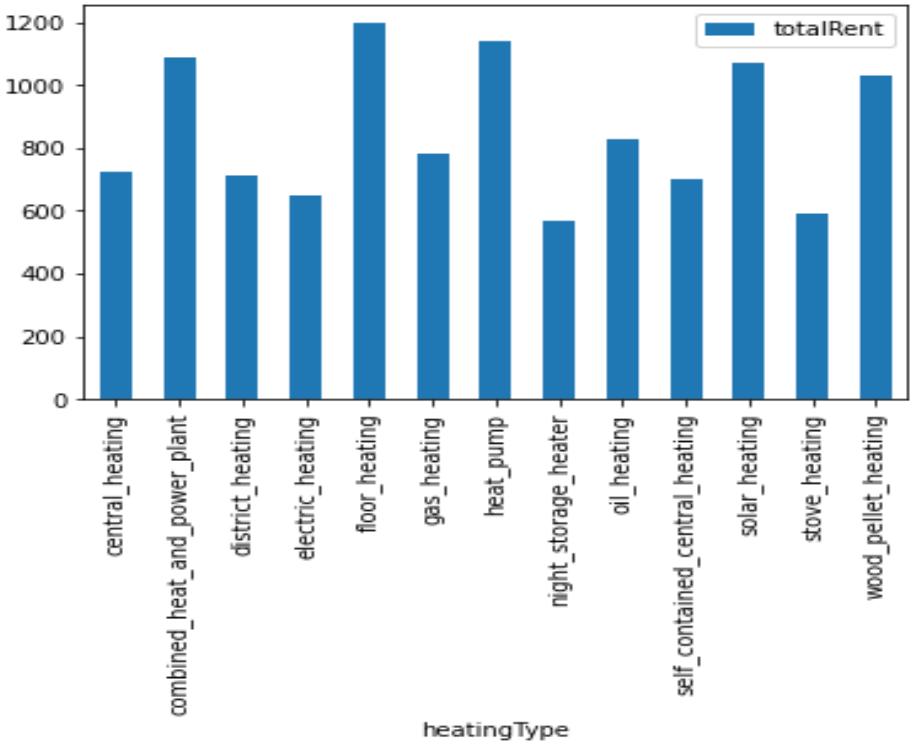
- در بین آگهی ها بیشترین نوع خانه مربوط به آپارتمان می باشد.



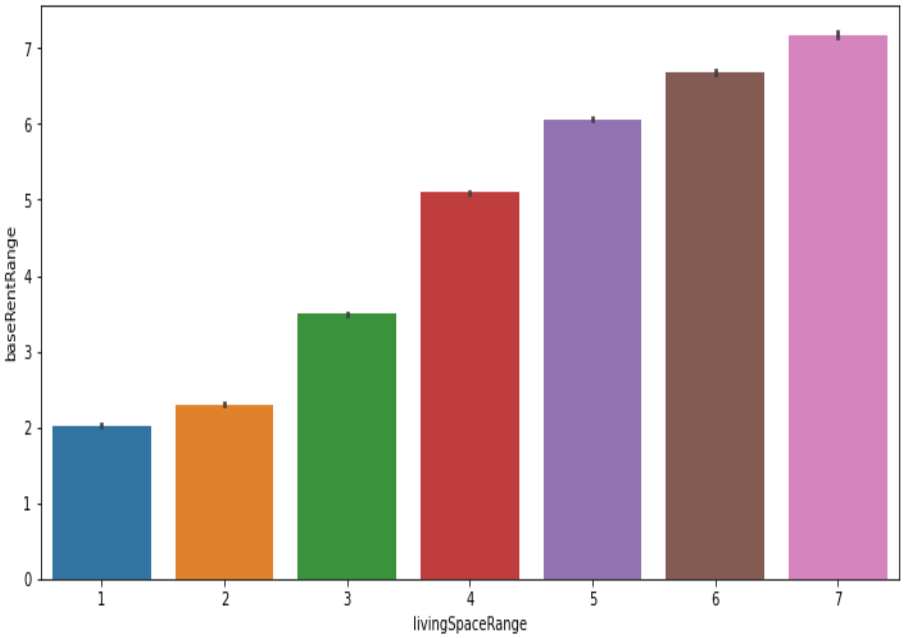
- در بین مناطق نوع 1 بیشترین اجاره کل مربوط به Hamburg و کمترین آن مربوط به Sachsen_Anhalt می باشد. میانگین مقدار اجاره کل بر اساس نوع منطقه نوع 1 به صورت زیر است:



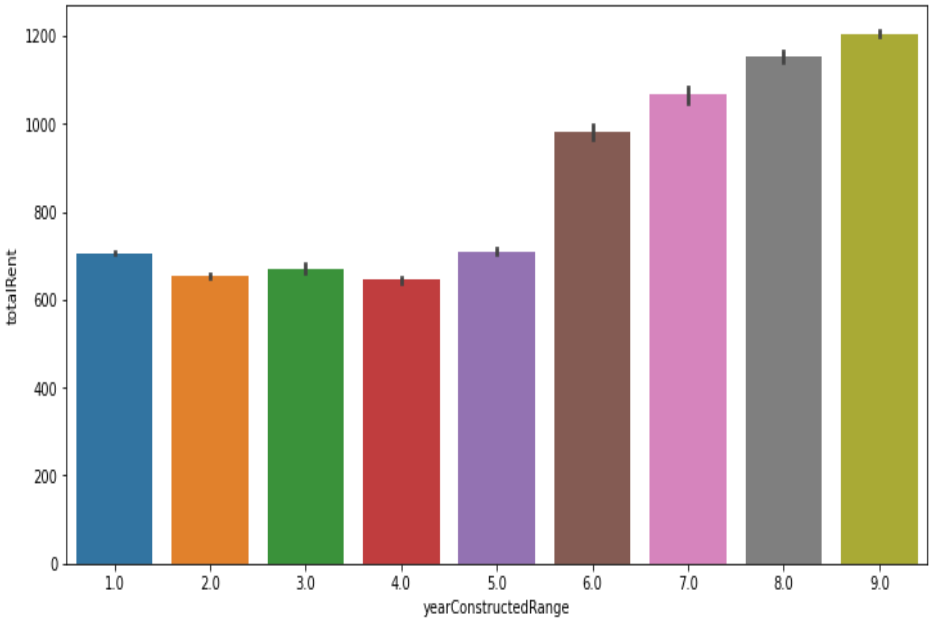
- در بین آگهی ها بیشترین اجاره کل مربوط به خانه هایی با گرمایش به صورت مرکزی central heating می باشد و کمترین آن مربوط به گرمایش به صورت night storage heater می باشد. میانگین مقدار اجاره کل بر اساس نوع گرمایش به صورت زیر است:



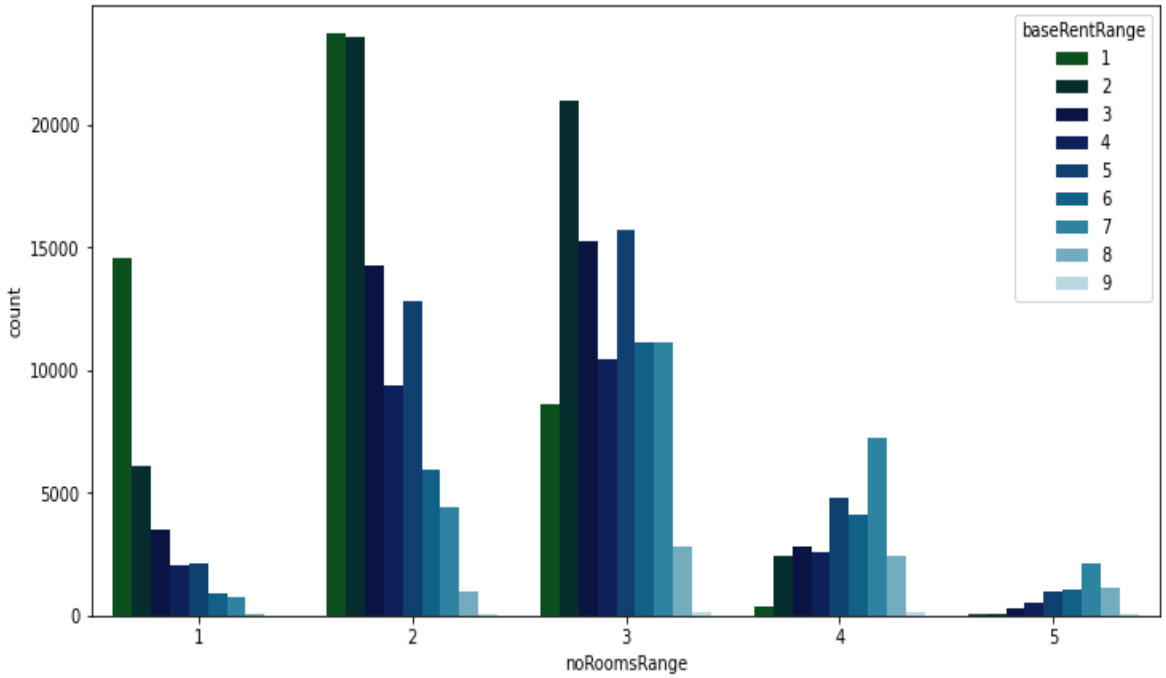
- با رشد میزان مساحت خانه ها، پایه اجاره ها نیز افزایش می یابد.



- دامنه سال ساخت ارتباط مستقیمی با میران اجاره کل دارد.



- خانه های با پایه اجاره 1 و 2 (در کل پایینتر) در بین آگهی های با تعداد اتاق کمتر فراوانی بیشتری دارند.



سوال 3-

داده های غیر عددی را به صورت ستون های دودویی تغییر یافتند و بعد با کمک رگرسیون بر روی میزان اجاره کل پیش بینی انجام شد.

سوال 4-

با کمک حلقه روی ستون ها در قسمت پر کردن داده های خالی و نیز پیدا کردن داده های پرت تغییر مشهودی در زمان اجرا مشاهده شد. ابتدا برای پر کردن داده های خالی در حالت بدون multiprocessing زمان صرف شده507 میلی ثانیه اما با multiprocessing 273میلی ثانیه بدست آمد.
برای پیدا کردن داده های پرت در حالت بدونmultiprocessing زمان صرف شده998 میلی ثانیه اما با multiprocessing584 میلی ثانیه بدست آمد.

سوال5-
برای قسمت پر کردن داده ها خالی با استفاده از dask کاهش زمان در حدود 100 میلی ثانیه مشاهده شد.