داده کاوی

پروژه اول

دیتاست دوم

دکتر فراهانی – دکتر شریفی

ابراهیم شکیباپور

۹۹۴۲۲۱۱۷

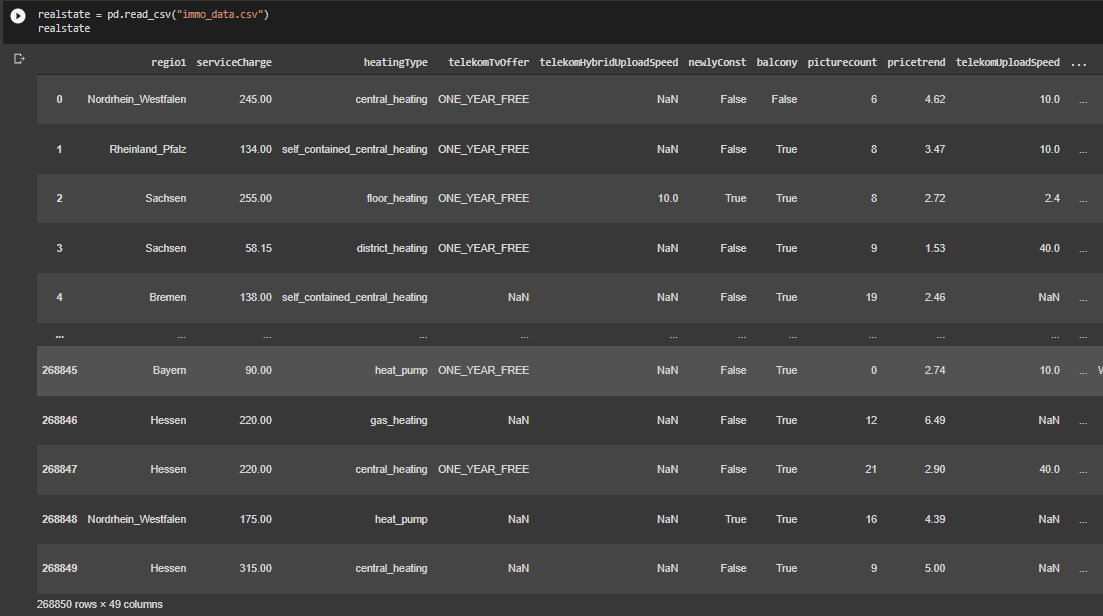
منبع داده

این داده ها از Immoscout24، بزرگترین پلتفرم املاک و مستغلات در آلمان استخراج شده است. Immoscout24 فهرست‌هایی برای املاک اجاره‌ای و خانه‌های فروشی دارد، با این حال، داده‌ها فقط شامل پیشنهادهایی برای املاک اجاره‌ای هستند.

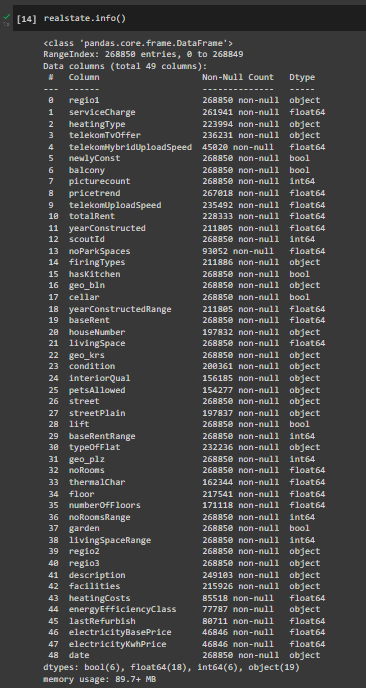
محتوا

مجموعه داده شامل بسیاری از ویژگی های مهم، مانند اندازه محل زندگی، اجاره، هر دو اجاره پایه و همچنین اجاره کل (در صورت وجود)، مکان (خیابان و شماره خانه، در صورت وجود، کد پستی و ایالت)، نوع انرژی و غیره. همچنین دارای دو متغیر حاوی توضیحات متنی مفصل تر است: توضیحات با متنی که پیشنهاد را توصیف می کند و امکاناتی که همه امکانات و وسایل رفاهی موجود را توصیف می کند، جدیدترین بازسازی و غیره. ستون تاریخ برای نشان دادن زمان خراش داده اضافه شد.

بررسی و تحلیل دیتاست

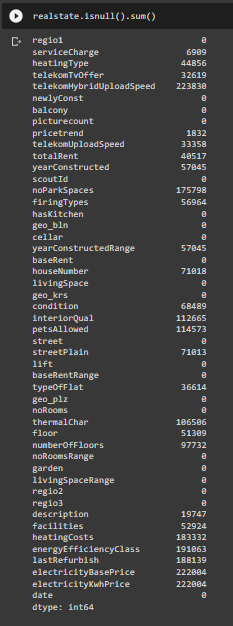


دیتاست موردنظر دارای ۲۶۸۸۵۰ سطر و ۴۹ ستون می باشد.

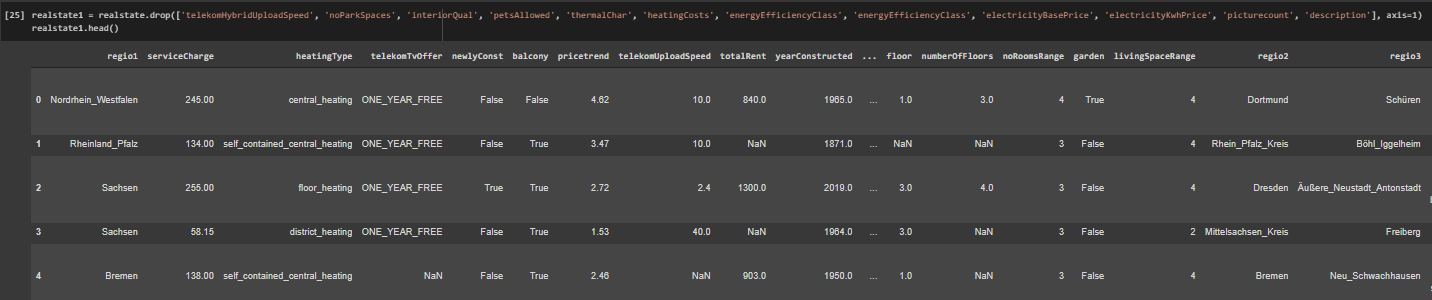


عناوین ستون ها و نوع داده متناظر با هر ستون را می توانید در بالا مشاهده کنید.

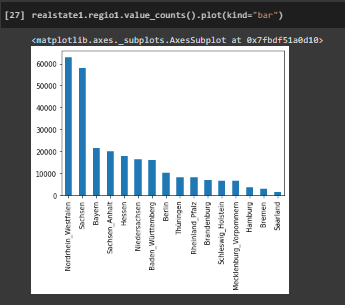
تعداد ۲۴ ستون کمی و ۲۵ ستون کیفی داریم.



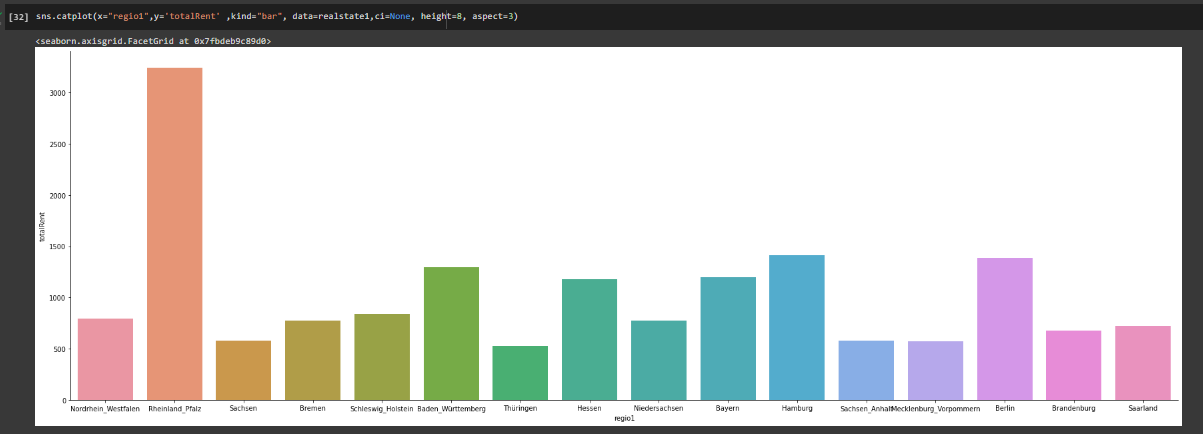
می بینیم در دیتاست ما ستون های بسیاری هستند که شامل کمبود و عدم حضور داده هستند. در بعضی از ستون ها تعداد داده های مفقود نزدیک به تعداد کل سطرها یا رکوردها می باشد. خیلی از این ستون ها باتوجه به سوال هایی که با آن مواجه هستیم از اهمیت چندانی برخوردار نیستند و می توانیم از آن ها چشم پوشی کرد. بعضی از ستون ها را حذف و بعضی از ستون ها را با مقادیر دلخواه جایگذاری می کنیم.



ستون هایی را که حاوی داده های مفقود زیادی بودند و یا فاقد ارزش محاسباتی بودند را حذف کردیم.



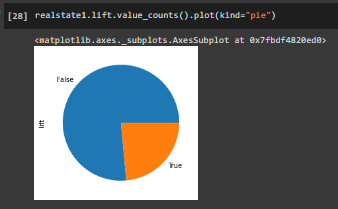
تعداد آگهی در مناطق مختلف



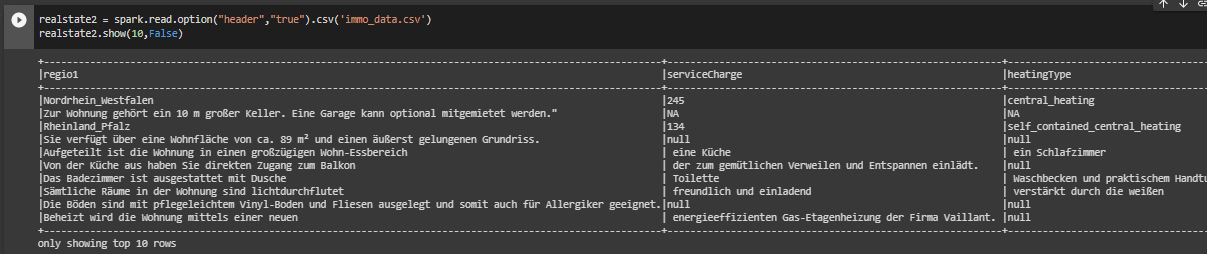
و همچنین بررسی قیمت ها در مناطق مختلف



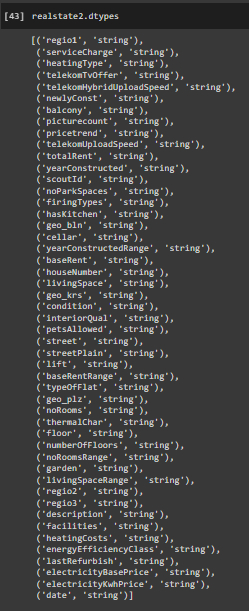
بررسی نسبت خانه هایی که بالکن دارند و خانه هایی که بالکن ندارند.



بررسی نسبت خانه هایی که آسانسور دارند و خانه هایی که آسانسور ندارند.



خواندن دیتاست با پای اسپارک



نمایش عناوین ستونها و نوع داده متناظر با آن ها