به نام خدا

گزارش مربوط به تمرین اول،دیتاست شماره 2

دیتاست مربوط به اجاره منازل در مناطق کشور آلمان

1) در گام اول دیتاست را دانلود و آن را با استفاده از گوگل درایو فراخوانی کردیم

2)در گام بعد با استفاده از کتابخانه پاندا و دستور read\_csv آن را خواندیم با توجه به این که فرمت فایل csv بود

3)با استفاده از توابع head,shape,info,dtypes اطلاعات کلی مربوط به دیتاست رو به دست آوردیم تا بتوانیم در گام های بعدی به کار ببریم و در کل با فیچر ها و نوع آن ها و یا تعداد کل رکورد ها آشنا شویم

4)با استفاده از تابع nunique متوجه شدیم که هر ستون یا فیچر دارای چه تعداد مقدار متمایز میباشد

5)با استفاده از دستور value\_countمقادیر هر فیچر و تعداد آن ها را به دست آوردیم

همچنین برای فیچر telekomUploadSpeed بااستفاده از دستورات آورده شده آن دسته ای که تعداد کمتر را داشتند در دسته otherقرار داده شد

6)در گام های بعدی بررسی رو فیچر های مختلف صورت گرفته همچنین فیچر telekomHybridUploadSpeed حذف شده چون یک لیست مرتب از اعداد بود و کارا به نظر نمیرسید

7)در گام های بعدی نیز با توجه به تعداد بالای فیچر ها آن دسته ای که به نظر ارتباط کمتری داشتند حذف میشود

8)با استفاده از رسم نودار به بررسی ارتباط بین مقدار پایه قیمت با منطقه جغرافیایی که به سه دسته تقسیم شده پرداختیم

9)در گام بعدی رنج قیمت پایه را محدود کردیم و در مرحله بعد به ارتباط قیمت پایه با heatingType پرداختیم

10)در گام بعد نیز با ترسم یم نمودار به بررسی قیمت خانه با سال ساخت آن پرداختیم

همین طور در ادامه بااستفاده از انواع مختلف نمودار ها به بررسی ارتباط قیمت با سایر فیچر ها مثل condition, interiorQual, typeOfFlat, noRooms پرداختیم

11)در گام بعدی نیز مطابق قبل داده ها نرمال شدند و مقادیر پرت شناسایی و حذف شدند

12)در گام بعدی به استفاده از multiprocess برروی تعریف تابعی که برای شناسایی و تغییر دادن فیچرهایی که object بودن پرداختیم که زمان اجرایی آن ها محاسبه شد

13)از Dask و pyspark نیز برای پاکسازی داده ها استفاده شد با اعمال آن ها بر تابع قبلی