به نام خدا

تمرین شماره ۱

مقدمه:

تمرین شماره ۱ در رابطه با داده هایی است که در رابطه با اجاره منازل در شهر نیویورک در سال ۲۰۱۹ است و از دیتاست شرکت Airbnb استفاده می شود. این تمرین شامل شش تسک است که در ادامه توضیحات لازم مربوط به انجام هر تسک ارایه شده است. همچنین در فایل مربوط به کدها در ؛ توضیحات لازم بصورت کامنت ارایه شده است.

تسک ۱:

با توجه به دیتاست داده شده ابتدا بعد از فراخوانی دیتاست ستون های id و host_id را حذف می کنیم زیرا این ستون ها فاقد اطلاعاتی هستند که در پردازش ما نقش دارند سپس ستون ها را بررسی می کنیم اگر ستونی بیشتر از ۵۰ درصد اطلاعاتش non باشد آن را حذف کرده و اگر این مقدار کمتر از ۵۰ درصد باشد مقادیر نا موجود را با میانگین اعداد همان ستون پر می کنیم همچنین ستون last_review را نیز حذف می کنیم زیرا هم مقادیر ناموجود آن بسیار است و همچنین نظر شخصی بنده بر آن است که این ستون در مقایسه با سایر پارا متر ها تاثیر خاصی بر پردازش ندارند. همچنین ستون هایی که مقادیر غیر عددی را دارند اگر حاوی مقدار نا موجود باشند با حداکثر فراوانی همان ستون پر می کنیم. در نهایت بر روی ستون neighborhood یک انکدر را ایجاد می کنیم. دلیل عدم ایجاد one_hot_encoder بر روی این ستون آن است که

به نام خدا

تعداد فیچرها با این اعمال بسیار افزایش می یابد. در نهایت بر روی دو فیچر neighborhood_qroup و room_type one_hot_encoder را اعمال میکنیم. سرانجام برای حذف داده های پرت از zscore استفاده میکنیم.

تسک ۲:

در این قسمت هم تعداد آگهی ها و فراوانی آن در هر منطقه جغرافیای ی راحساب کرده و رسم میکنیم. همچنین تاثیر هر یک از پارامتر ها بر قیمت را بررسی کرده و رسم میکنیم .
برای مثال رابطه بین latitude و قیمت رابطه ای خطی نیست و می توان فهمید که فراوانی خانه ها در محدود ۴۰.۷ تا ۴.۷۵ بیشتر است و در قیمت های بالا هم به همین ترتیب است.

تسک ۳ و ۴:

جواب با توجه به گزارش فایل ژوپیتر مشخص هستند

تسک ۵:

در ابتدا از ضریب پیرسون استفاده کردیم که جز correlation test می باشد و نشان دهنده مقدار هم بستگی بین فیچر هست. در دو مرحله ی بعد از آزمون های students t_test paired students t_test. که جزو آزمون فرضیه های آماری پارامتریک هستند استفاده شد که از آنها برای مقایسه نمونه داده ها استفاده می شود و نشان دهنده این است که آیا میانگین دو نمونه مستقل به طور قابل توجهی متفاوتند یا خیر.

به نام خدا

سرانجام در دو مرحله بعد از آزمون های mann whitney وKruskal استفاده میکنیم که جزو آزمون فرضیه های آماری ناپارمتریک هستند و نشان دهنده این هستند که دو نمونه توزیع یکسانی دارند یا خیر.

تسک ۶:

در نهایت این پردازش ها دیتاست را به دو مقدار آموزش و تست جدا کرده **(**80 درصد داده های آموزشی اند**)** و مدل را آموزش میدهیم و مشاهده میکنیم که دقت یادگیری مدل با داده های آموزشی ۳۴ درصد است و سپس با داده های تست مدل را تست میکنیم و متوجه میشویم که دقت پیشبینی مدل ۳۵ درصد است.