به نام خدا

گزارش مربوط به تمرین اول،دیتاست شماره 1

دیتاست مربوط به اجاره منازل در شهر نیویورک

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 فیچرneighbourhood بااستفاده از دستورات آورده شده چون تعداد مقادیر زیاد بود آن دسته ای که تعداد کمتر از 30 را داشتند در دسته otherقرار داده شد

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

7)با استفاده از تابع describeاطلاعات عددی مربوط به فیچر های عددی رو بررسی کردیم مثل میانه یا ماکسیمم

8)در گام بعدی داده های null رو بررسی کردیم شامل تعداد آن ها و اینکه در کدام یک از فیچر ها تعداد بیشتری دارند

9)در گام های بعدی ستون هایی که مقادیر بیشتری nullداشتند را با میانه و مد آن مقادیر پر کردیم و یا با unknown

10)در گام بعد داده های پرت را شناسایی کردیم با استفاده از به دست آوردن upper\_range, lower\_range و حذف کردندادههای خارج از این رنج

11)داده هایی که فرمت objectدارند را شناسایی و با استفاده از dummies\_featureآن ها را تبدیل کردیم که در واقع به ازای هر متغیر هر ستون، یک ستون به داده ها اضافه میکند

12)با استفاده از روش PCA اقدام به کاهش تعداد ستون ها کردیم

13)با استفاده از ترسیم نمودار اقدام به بررسی قیمت اجاره در مقایسه با فیچر های دیگر کردیم و همین طور برای فیچرهای دیگر

14)با استفاده از groupbyفیچر مربوط به قیمت و محله های آورده شده را کنار هم مقایسه کردیم و در ادامه برای فیچر های دیگر نیز این کار را ادامه دادیم

15)با استفاده از کتابخانه matplotlibوseabornاقدام به رسم نمودار های مختلف کردیم که قیمت خانه را با فیچر های دیگر مثل تعدادتبلیغات یا محله آن مقایسه میکرد

مثلا متوجه شدیم که در محله manhattanقیمتخانه در مقایسه با محله های دیگر بیشتر است

16)با استفاده از دستورات آورده شده 10 صاحب خانه ای که بیشترین فعالیت یا آگهی را داشتند را شناسایی کردیم و همچنین اطلاعات مربوط به خانه هایی که گزارش شده بود

17)توزیع یا پراکندگی مربوط به قیمت خانه ها را با رسم نمودار بررسی کردیم

18)درگام بعد بااستفاده از کتابخانه و نمودارplot,scatter پراکندگی جغرافیایی خانه ها را بررسی کردیم ،این که در چه مناقی بیشتر یا کمتر هستند و همچنین شعاع هر نقطه هر چه بیشتر باشد به معنی قیمت بیشتر آن نقطه میباشد

در نمودار بعدی تعداد آگهی ها نیز واردشده که با تغییرات رنگ تاثیر و فراوانی آن ها گزارش میشود

19)آزمونهای فرض آورده شده که شامل paired sample T-test,Z-testوکای2میباشدکه احتمال وابستگی قیمت خانه ها با محله یا تعداد آگهی ها و یا احتمال برابری میانگین حدس زده شده با میانگین واقعی آورده شده

20)در مرحله بعد داده ها به فرم استاندارد در آورده شدن که باید در مراحل قبلی صورت میگرفت ولی نمودار ها رو خراب میکرد

21)یک مدل رگرسیون خطی پیاده سازی شده که با ارور همراه شده ولی هیچ فیچر object یا boolنداریم