به نام خدا

گزارش مربوط به Germany rental

زهرا احمدنژاد

شماره دانشجویی: 400422011

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

در اینجا نیز اول کتابخانه هایی که لازم و با آنها کار داریم را ران میکنیم.

from google.colab import files

files.upload()

!mkdir -p ~/.kaggle

!cp kaggle.json ~/.kaggle/

!chmod 600 ~/.kaggle/kaggle.json

!kaggle datasets download -d corrieaar/apartment-rental-offers-in-germany

import zipfile

zip\_ref = zipfile.ZipFile('apartment-rental-offers-in-germany.zip', 'r')

zip\_ref.extractall('files')

zip\_ref.close()

!unzip \\*.zip  && rm \*.zip

بعد دیتاست موردنظر را دانلود کرده و در کولب ایمپورت میکنیم تا بتوانیم از آن استفاده کنیم.

immo\_data.info()

این تابع اطلاعات مربوط به این دیتاست را به ما نشان میدهد. مانند تعداد ستون، نوع ستون، اسم ستون، نوع دیتا و...

#checking type of every column in the dataset

immo\_data.dtypes

بعد نوع متغیر هر ستون را بررسی میکنیم.

immo\_data.isnull().sum()

با استفاده ازین داده میخواهیم ببینیم در چه ستون هایی نال و مقادیر از دست رفته وجود دارد و مقدار آنها چقدر است.

def missing\_values(immo\_data,norows):   # input by the immo\_data and the number of rows that you want to show

    total = immo\_data.isnull().sum().sort\_values(ascending=False)

    percent = ((immo\_data.isnull().sum().sort\_values(ascending=False)/immo\_data.shape[0])\*100).sort\_values(ascending=False)

    missing\_data = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])

    return(missing\_data.head(norows))

میخواهیم ببینیم در هر ستون چند درصد از داده های آن نال است. چون تعداد نال ها در ستون ها زیاد است و به همین دلیل بررسی میکنیم ببینیم در هر ستون چند درصد نال وجود دارد.

چون تعداد نال ها زیاد است در اینجا به ازای هر ستون تعداد نال ها را شمرده و بر اساس زیاد به کم مرتب کرده و در ستون percent تعداد نال ها را تقسیم بر تعداد کل هر ستون کرده است.

missing\_data = missing\_values(immo\_data,20)

# drop the data where the columns contains more than 30%

immo\_data = immo\_data.drop((missing\_data[missing\_data['Percent'] > 30]).index,1)

در اینجا توضیح میدهیم که چون تعداد نال ها زیاد است و من نمیتونم تصمیم بگیرم که با همشون چیکار کنم میام هر ستونی که تعداد نال آن بیشتر از30 درصد است را دراپ میکنم. و برام مهم نیست جیه جون تعداد نال هاش خیلی زیاده به درد من نمیخوره.

immo\_data.dropna(subset=['totalRent'],inplace=True)

و مهم تر از آن این است که من چون میخواهم روی رنت و اجاره صحبت کنم ستون total rent برام خیلی مهم است پس این ستون را حتما نگه میدارم و نمیندازمش ولی هر سطری که total rent آن نال باشد هم به درد من نمیخورد و نمیتونم نظری درباره اجارش بدهم.