Germany Rental Offers

در این تمرین 268850 ریکورد از اطلاعات خانه های اجاره ای آلمان در اختیار داریم که دارای 49 فیچر می باشند.
ابتدا داده های null را بررسی می کنیم. در فیچر numberOffloors تعداد 97732 داده null داریم که آن ها را به کمک مقدار فیچر floor پر می کنیم و بعد از این کار تعداد داده های null به 39942 کاهش می یابد.
در فیچر totalRent به تعداد 40517 null داریم که با توجه به تعریف این فیچر، با جمع کردن مقادیر فیچر های baserent, heatingCosts, serviceCharge آن را بدست می آوریم و بعد از این کار تعداد null ها به 28235 کاهش می یابد.
در زیر درصد داده های null در هر ستون را مشاهده می کنیم:

regio1	0.000000
serviceCharge	2.569834
heatingType	16.684397
telekomTvOffer	12.132788
telekomHybridUploadSpeed	83.254603
newlyConst	0.000000
balcony	0.000000
picturecount	0.000000
pricetrend	0.681421
telekomUploadSpeed	12.407662
totalRent	10.502139
yearConstructed	21.218151
scoutId	0.000000
noParkSpaces	65.388879
firingTypes	21.188023
hasKitchen	0.000000
geo_bln	0.000000
cellar	0.000000
yearConstructedRange	21.218151
baseRent	0.000000
houseNumber	26.415473
livingSpace	0.000000

geo_krs	0.000000
condition	25.474800
interiorQual	41.906267
petsAllowed	42.615957
street	0.000000
streetPlain	26.413614
lift	0.000000
baseRentRange	0.000000
typeOfFlat	13.618747
geo_plz	0.000000
noRooms	0.000000
thermalChar	39.615399
floor	19.084620
numberOfFloors	14.856611
noRoomsRange	0.000000
garden	0.000000
livingSpaceRange	0.000000
regio2	0.000000
regio3	0.000000
description	7.344988
facilities	19.685326
heatingCosts	68.191185

energyEfficiencyClass	71.066766
lastRefurbish	69.979171
electricityBasePrice	82.575414
electricityKwhPrice	82.575414
date	0.000000

فیچر هایی که بیشتر از 40% داده null دارند و همچنین برخی از فیچر هایی که به نظر به کار نمی آیند را حذف می کنیم که در زیر لیست این فیچر ها آورده شده است:

```
'picturecount',
'scoutId',
'yearConstructedRange',
'yearConstructed',
'thermalChar',
'houseNumber',
'telekomHybridUploadSpeed',
'noParkSpaces',
'interiorQual',
'petsAllowed',
'heatingCosts',
'energyEfficiencyClass',
'lastRefurbish',
'electricityBasePrice',
'electricityKwhPrice']
```

```
'regio2',
'regio3',
'geo_bln',
'description',
'geo_krs',
'street',
'streetPlain',
'geo_plz',
'noRoomsRange',
'livingSpaceRange',
'condition',
'date',
'firingTypes',
'facilities',
'yearConstructed',
'baseRentRange',
'noRoomsRange',
```

بعد از این کار تعداد فیچر ها از 49 تا به 19 تا کاهش می یابد. حال برخی از تناقض های منطقی موجود در ریکورد ها را بررسی می کنیم مثل:

1- مقدار serviceCharge منفی نباشد.

2- اگر مقدار numberOfFloor نامنفی باشد، نباید اندازه floor بیشتر از numberOfFloor باشد.

3- اگر مقدار فیچر balcony برابر true است، نباید مقدار floor منفی باشد.

4- از آنجایی که فیچر totalRent ستون هدف است، نباید null باشد.

ریکورد هایی که شروط بالا را نداشته باشند، از دیتاست حذف می شوند که این کار باعث می شود تقریبا 30000 ریکورد کم شود.

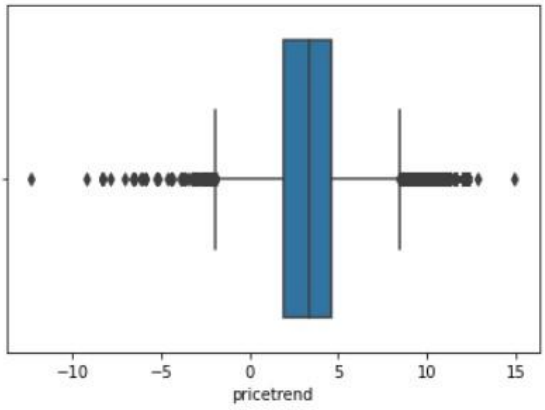
مقادیر null فیچر های numberOfFloor, pricetrend, serviceCharge را با میانگین این فیچر ها پر می کنیم.

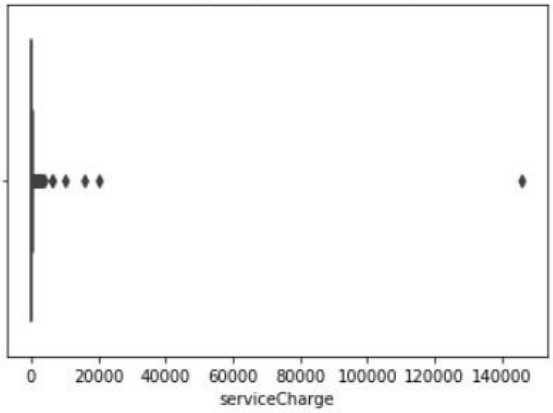
مقادیر null فیچر floor را با عدد یک و فیچر telekomUploadSpeed را با عدد 0 پر می کنیم.

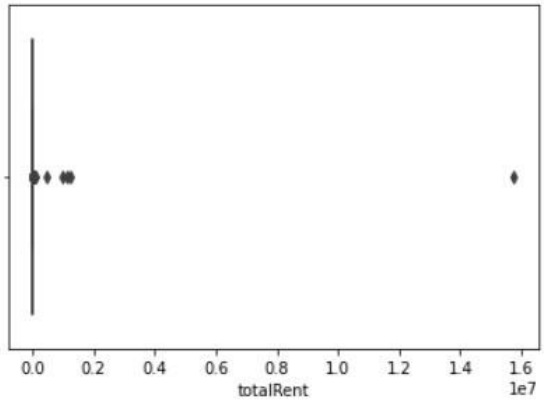
مقادیر null فیچر های کتگوریکتال heatingType, telekomTvOffer, typeOfFloat را با مد این فیچر ها پر می کنیم.

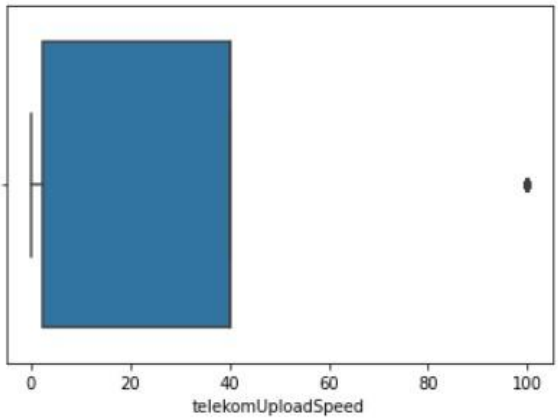
حال دیتاست دیگر داده null ندارد.

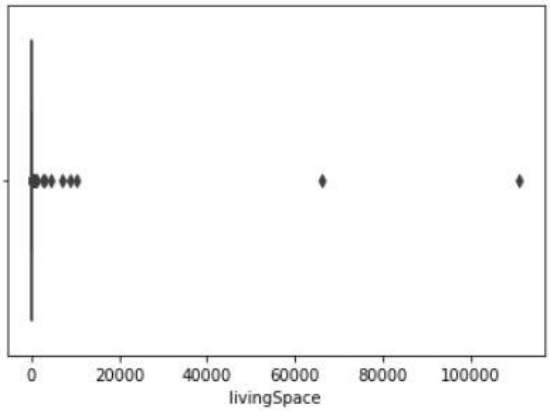
در زیر نمودار جعبه ای مربوط به برخی از فیچر ها را می بینیم که به وضوح دارای داده های پرت هستند. (نام هر فیچر در زیر نمودار آن ذکر شده)

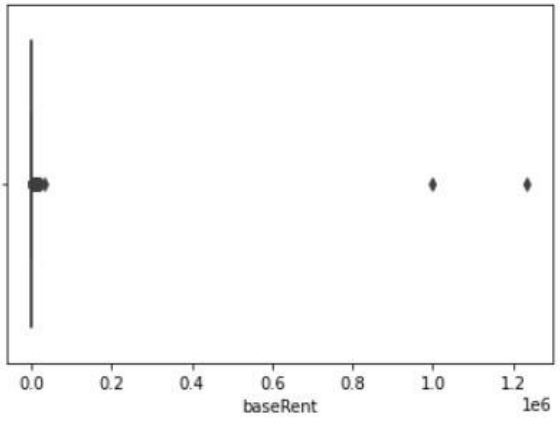


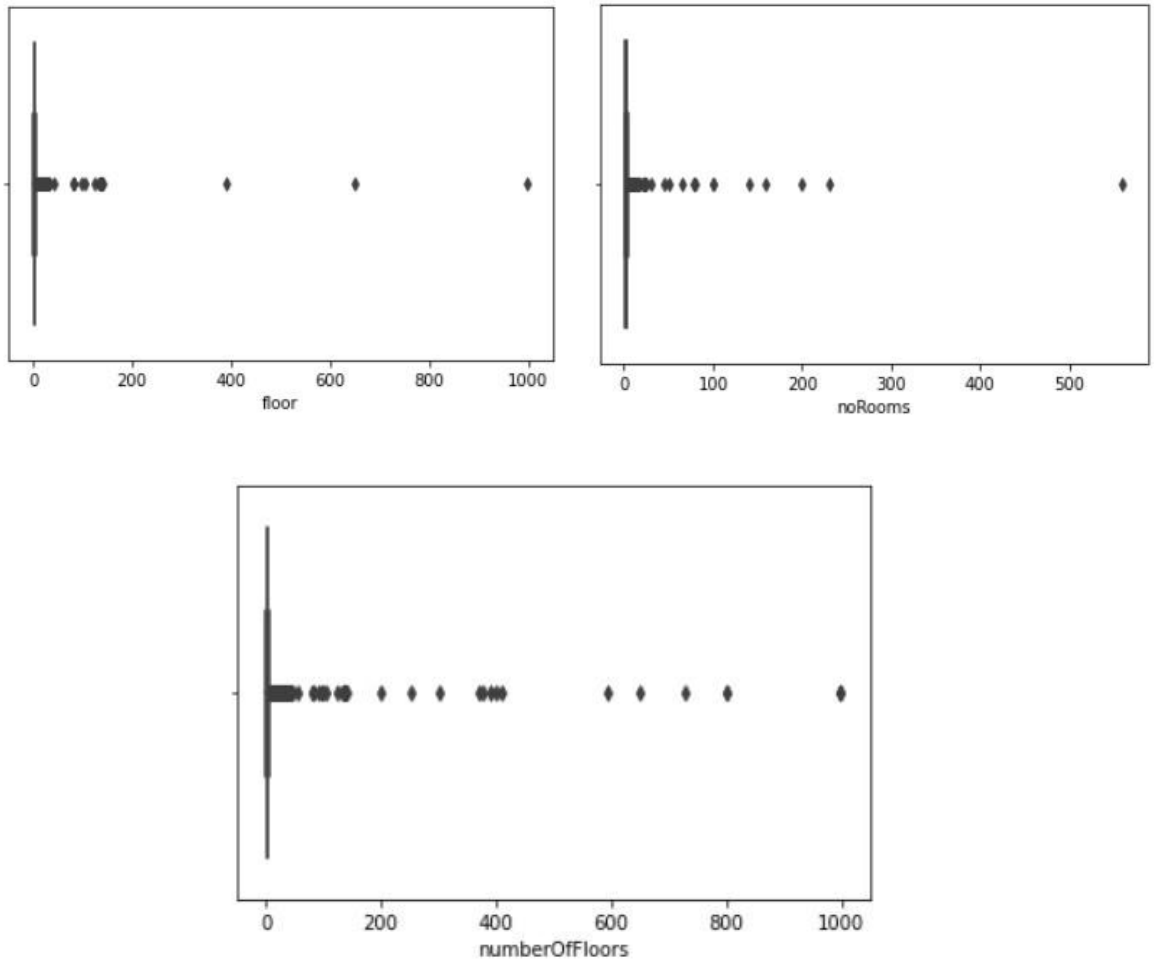












تابع del_outliers داده ه ای پرت این فیچر ها را با کمک روش IQR حذف می کند. یک بار این تابع را به صورت multi process با 2 تا cpu اجرا می کنیم. مقدار runtime در این حالت برابر زیر است:

```
run time : 0.9017238616943359
```

حال همین تابع را به صورت معمولی اجرا می کنیم و runtime این حالت برابر است با:

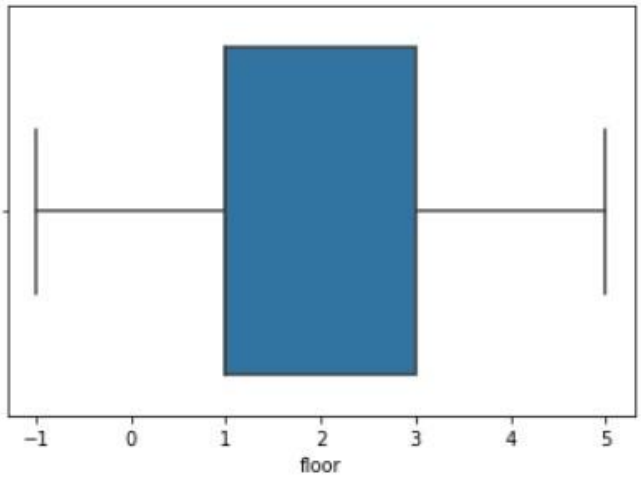
```
run time : 0.5075664520263672
```

افزایش مقدار runtime در حالت multi process احتمالا به دلیل زمان بر بودن فرایند ساخت process می باشد.

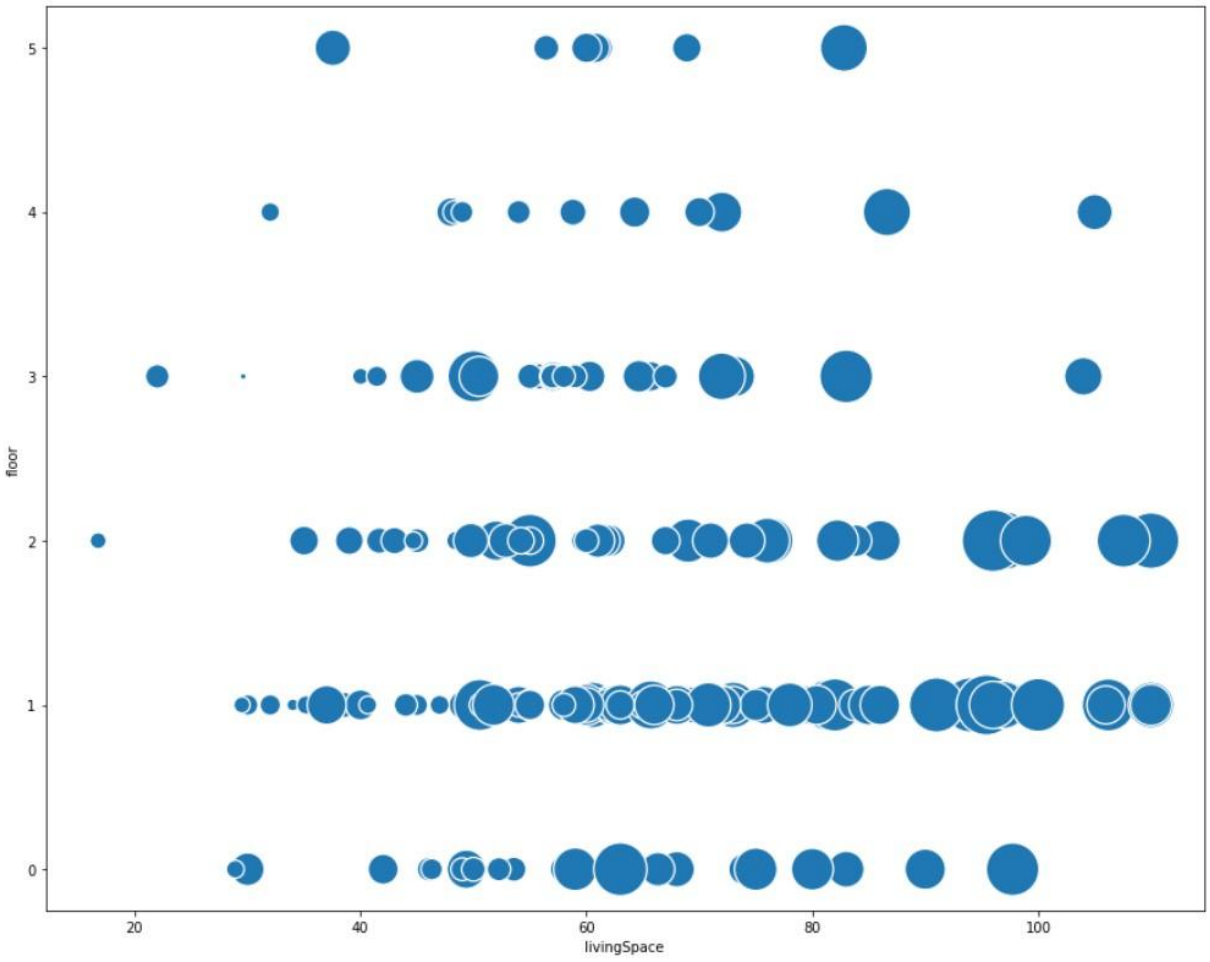
خروجی این تابع به صورت زیر است که در نهایت 196490 ریکورد باقی می میاند:

```
value of IQR of column serviceCharge ->86.0
Shape before delete outliers of column serviceCharge: (239267, 19)
New Shape  after delete outliers of column serviceCharge: (229812, 19)
------------------------
value of IQR of column pricetrend ->2.58
Shape before delete outliers of column pricetrend: (229812, 19)
New Shape  after delete outliers of column pricetrend: (225726, 19)
------------------------
value of IQR of column telekomUploadSpeed ->37.6
Shape before delete outliers of column telekomUploadSpeed: (225726, 19)
New Shape  after delete outliers of column telekomUploadSpeed: (225596, 19)
------------------------
value of IQR of column totalRent ->449.0
Shape before delete outliers of column totalRent: (225596, 19)
New Shape  after delete outliers of column totalRent: (215505, 19)
------------------------
value of IQR of column baseRent ->360.0
Shape before delete outliers of column baseRent: (215505, 19)
New Shape  after delete outliers of column baseRent: (211021, 19)
------------------------
value of IQR of column livingSpace ->28.0
Shape before delete outliers of column livingSpace: (211021, 19)
New Shape  after delete outliers of column livingSpace: (206011, 19)
------------------------
value of IQR of column noRooms ->1.0
Shape before delete outliers of column noRooms: (206011, 19)
New Shape  after delete outliers of column noRooms: (203843, 19)
------------------------
value of IQR of column floor ->2.0
Shape before delete outliers of column floor: (203843, 19)
New Shape  after delete outliers of column floor: (199708, 19)
------------------------
value of IQR of column numberOfFloors ->2.0
Shape before delete outliers of column numberOfFloors: (199708, 19)
New Shape  after delete outliers of column numberOfFloors: (196490, 19)
```

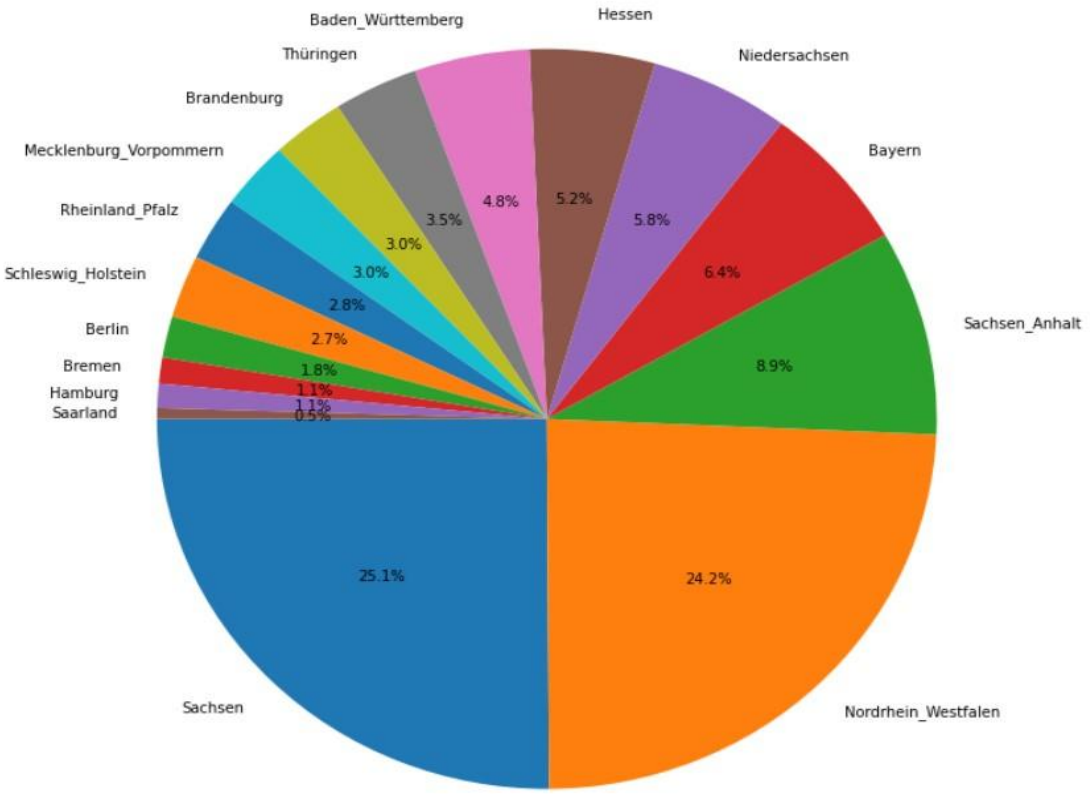
شکل زیر نمودار جعبه ای مربوط به فیچر floor را بعد از حذف داده های پرت نشان می دهد:



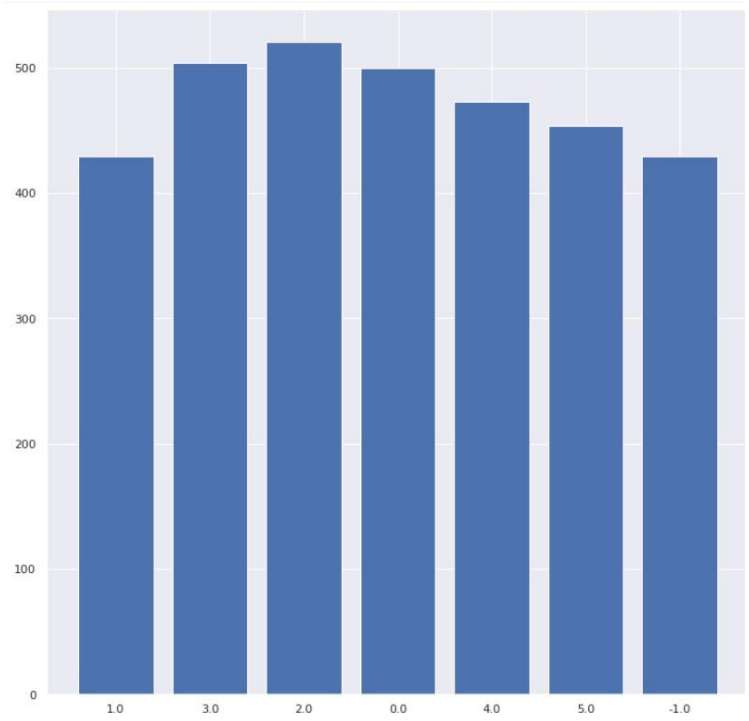
در شکل زیر محور عمودی بیانگر floor و محور افقی بیانگر livingSpace هستند و اندازه دایره هایی که داده ها را نشان می دهد با مقدار totalRent آن داده ها ارتباط مستقیم دارد.(این نمودار برای یک سمپل شامل 200 ریکورد رسم شده است)



شکل زیر بیانگر نحوه توزیع این داده ها در مناطق مختلف است:



در نمودار زیر مقدار میانگین baseRent در مقادیر مختلف ستون floor نمایش داده شده است:



در فیچر های دیتاست که دارای مقادیر Bool هستند، به جای true مقدار 1 و به جای false مقدار 0 را قرار می دهیم.
از آنجایی که ستون telekomTvOffer دارای 3 مقدار منحصر به فرد است، روی این فیچر label encoding را اجرا می کنیم.
ستون های regio1, heatingType, typeOfFlat دارای مقادیر منحصر به فرد زیادی هستند لذا روی این ستون ها target encoding را اجرا می کنیم.

امیرحسین باباجانی
97222009