به نام خدا

محمدعلی رشادی

گزارش تمرین سری ۱

دیتاست اول

۱.

خوشبختانه دیتای از دست رفته و تکراری در دیتاست موجود نبود

```
df.isnull().sum()
```

```
battery_power    0
blue             0
clock_speed      0
dual_sim         0
fc               0
four_g           0
int_memory       0
m_dep            0
mobile_wt        0
n_cores          0
pc               0
px_height        0
px_width         0
ram              0
sc_h             0
sc_w             0
talk_time        0
three_g          0
touch_screen     0
wifi             0
price_range      0
dtype: int64
```

```
df.duplicated().sum()
```

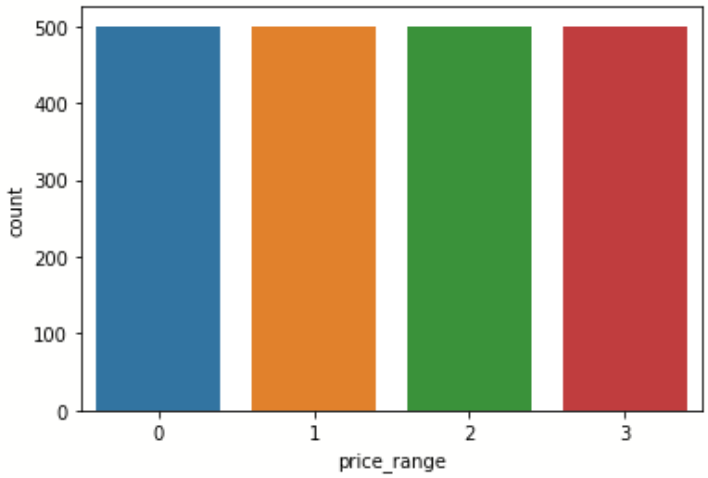
```
0
```

با بررسی مشخصات ویزگی ها و چارک های هر ویژگی توسط تابع `describe` دریافتیم که داده پرت خاصی وجود ندارد و دیتاست تمیز است.

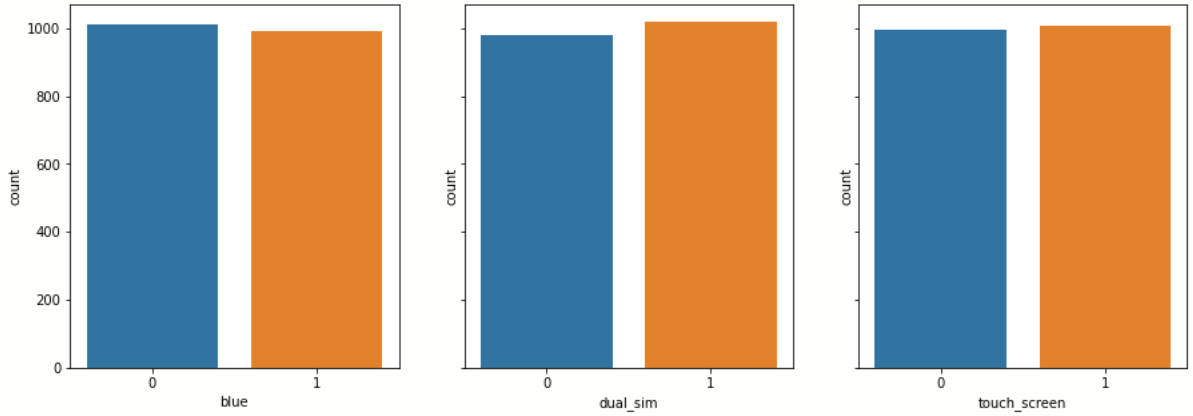
- همچنین در این بخش از `dask` به عنوان بخش امتیازی استفاده شده است.

.۲

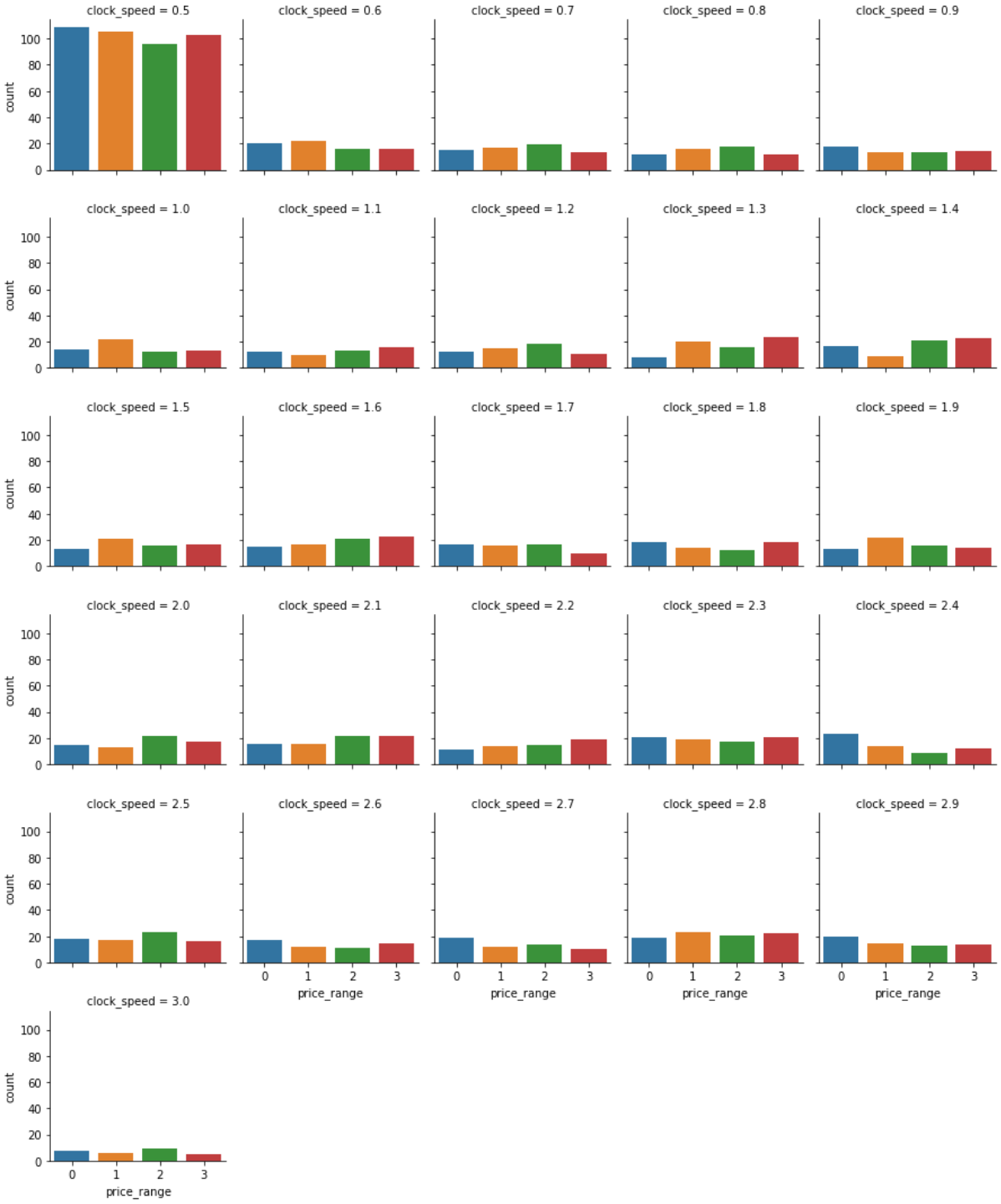
می بینیم که از قیمت های مختلف تقریبا به مقدار یکسان دیتا موجود است و دیتاست متوازن است.



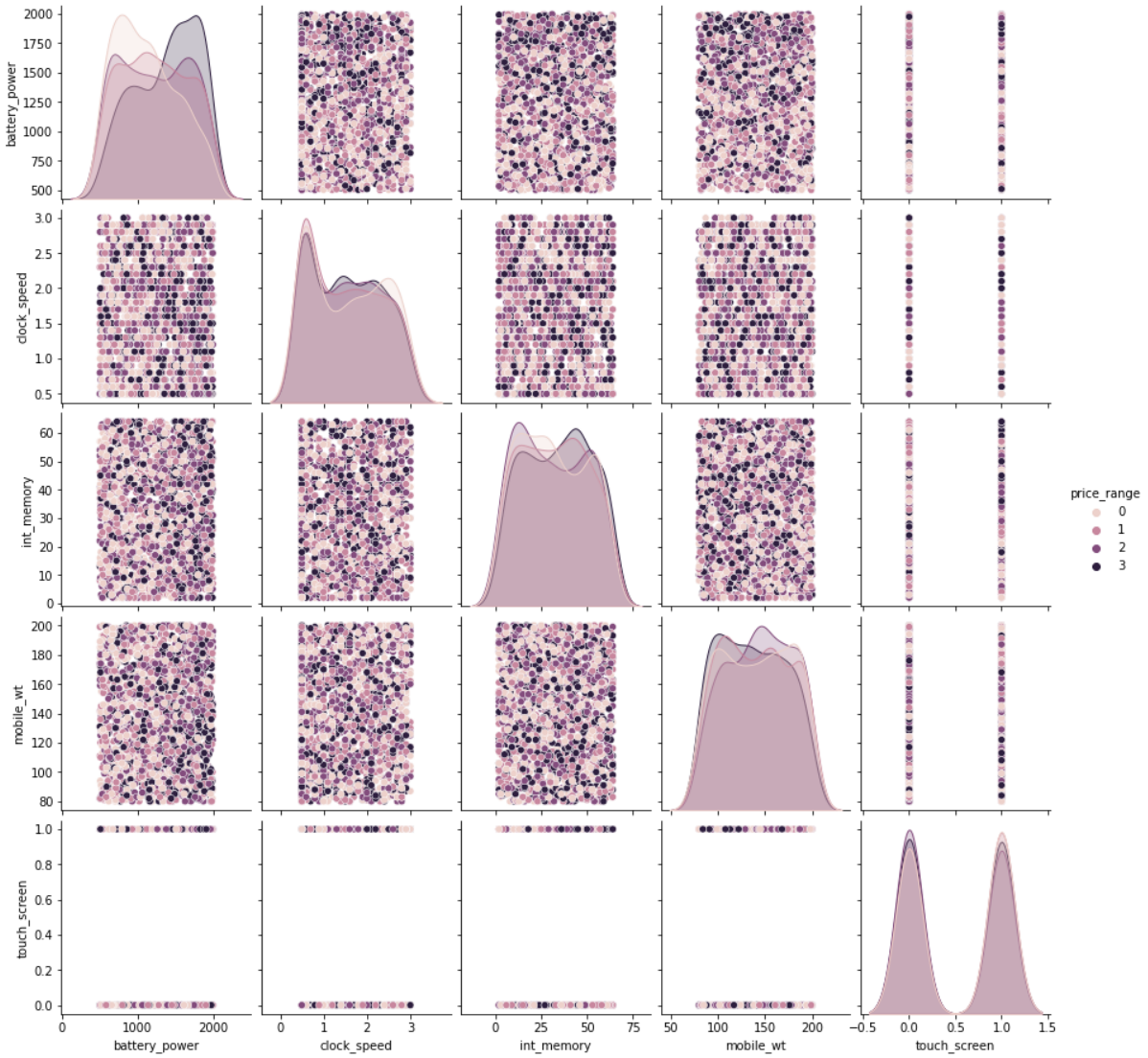
همچنین مقدار ویزگی های بولین را در تصویر زیر میبینیم که نشان دهنده توازن از انواع مختلف است



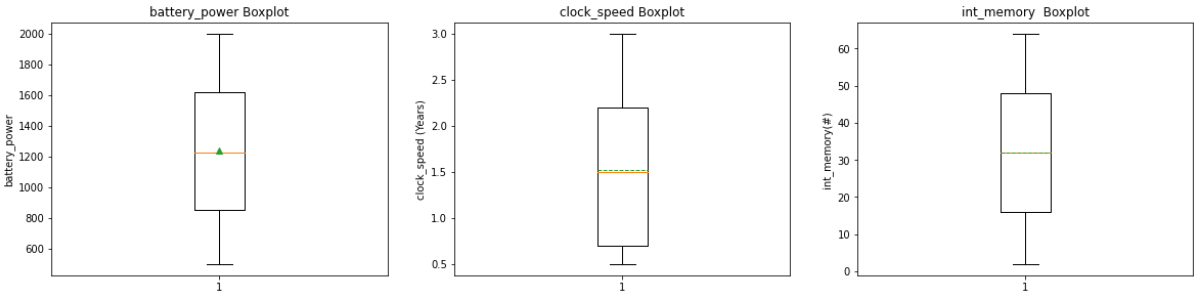
clock_speed های مختلف را در تصویر زیر شاهدیم که بیشترین فراوانی برای مقدار ۰٫۵ است.



در تصویر زیر نمودار جفتی را شاهد هستیم که ارتباط ویژگی ها را با یکدیگر مشخص کرده که بعضی پراکنده و بعضی دارای ارتباط هستند. (برای درک بهتر نمودار از جدول همبستگی نیز در پایین اسفتاده شده است.

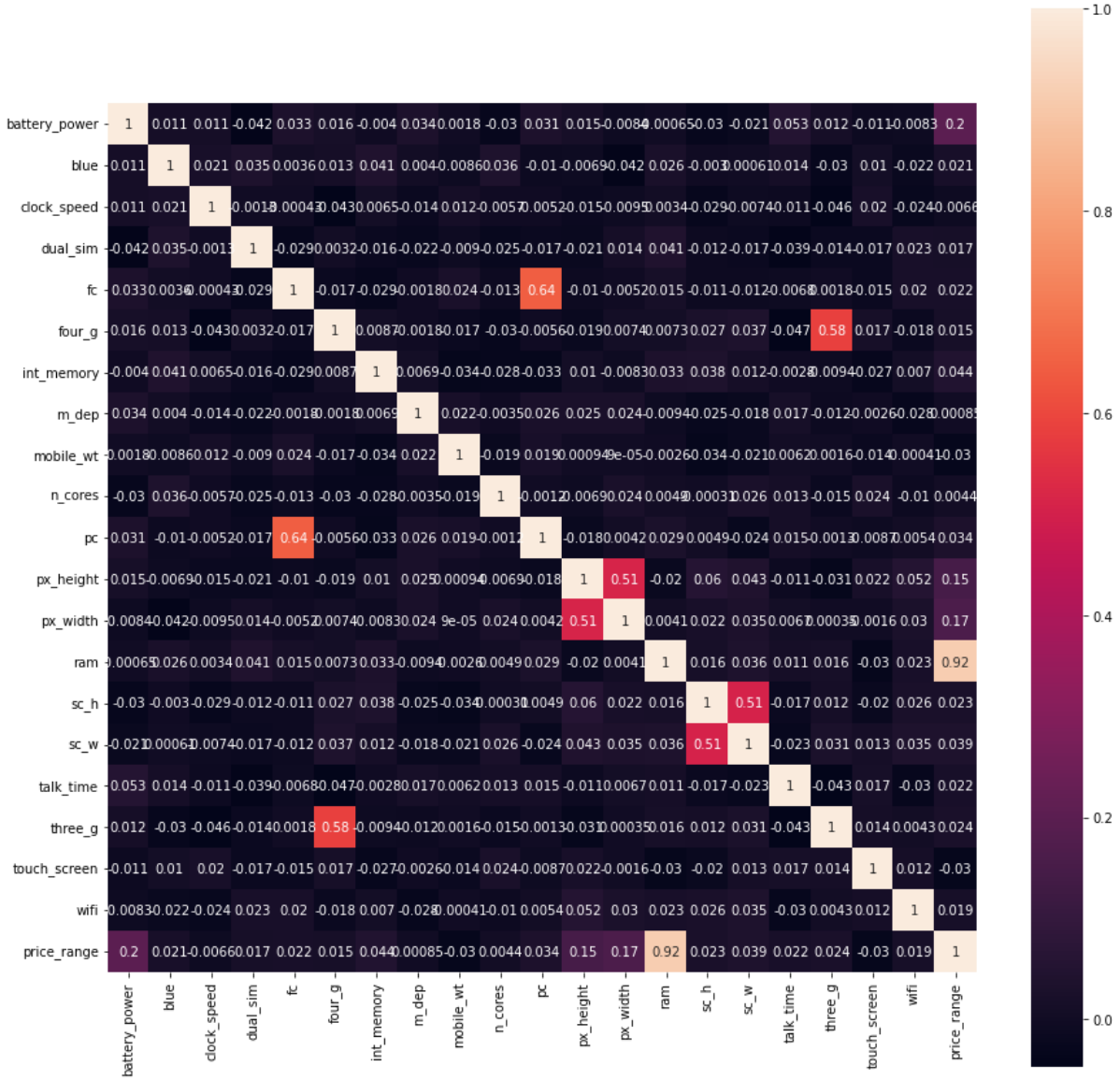


همچنین باکس پلات ویژگی ها عددی را نیز به تصویر کشیدیم که طبق صورت سوال یک داده پرتی نیز شاهد نیستیم



جدول همبستگی نیز همبستگی جفتی ویژگی ها را نیز نمایش می دهد

برای مثال مقدار رم با قیمت همبستگی زیادی دارد.



.۳

از پنج آزمون فرض متفاوت بهره بردیم

۱. افراد ۶۰ دلار برای خرید گوشی میپردازند که فرضیه رد شد. (ttest1.samp)

t stat : -2339.4149268567126 , p_value : 0.0

reject null hypothesis

۲. دو سیم کارت بودن گوشی در قیمت موثر است که فرضیه مورد قبول واقع شد. (ttest.ind)

t stat : 0.839810370538735 , p_value : 0.4011151943555157

accept null hypothesis

۳. وای فای داشتن گوشی در قیمت موثر است که مورد قوبل واقع شد (f-oneway)

f stat : 0.6081961720936766 , p_value : 0.4355601630197592
accept null hypothesis

.۴

بین مصرف باتری و صفحه نمایش لمسی ارتباط وجود دارد که رد شد(contingency)

Retain H0,There is no relationship between 2 categorical variables

۵. بین عرض و طول صفحه نمایش ارتباط وجود دارد که رد شد.(ttest_rel)

t stat : -62.5719510595071 , p_value : 0.0

reject null hypothesis

۴.

از سه مدل svm ، knn و decision tree بهره بردیم.

پارامتر های مورد بررسی svc

ضریب جریمه C – هسته kernel -درجه بردار degree

پارامتر های مورد بررسی KNN

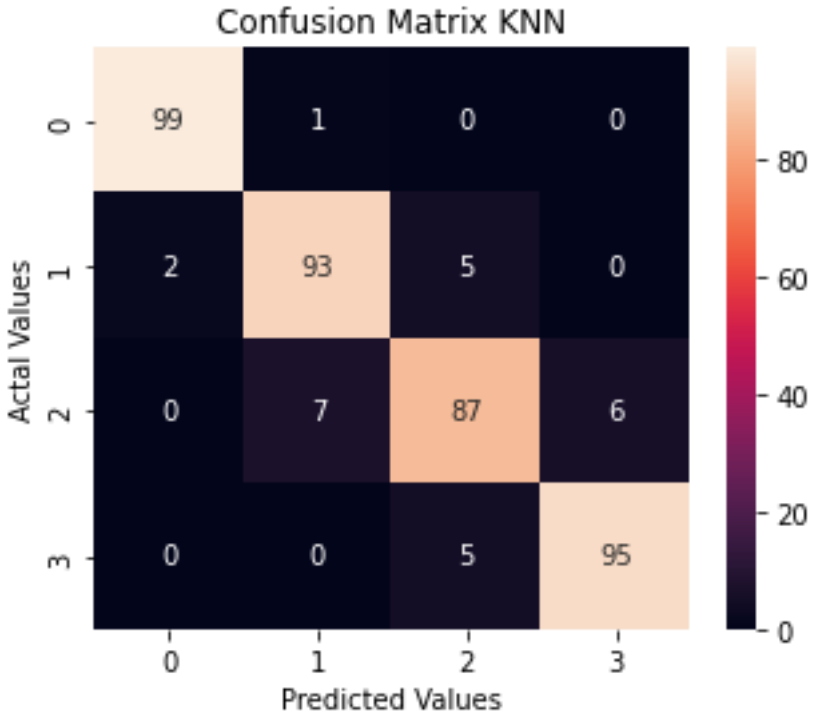
تعداد همسایه n_neighbors– متریک فاصله metric-توان متریک p

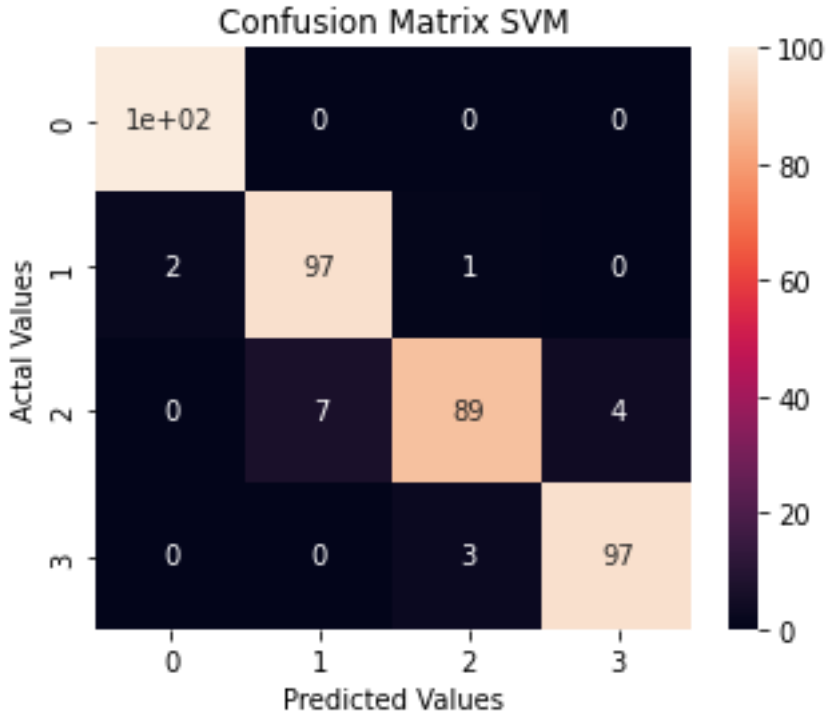
پارامتر های مورد بررسی DT

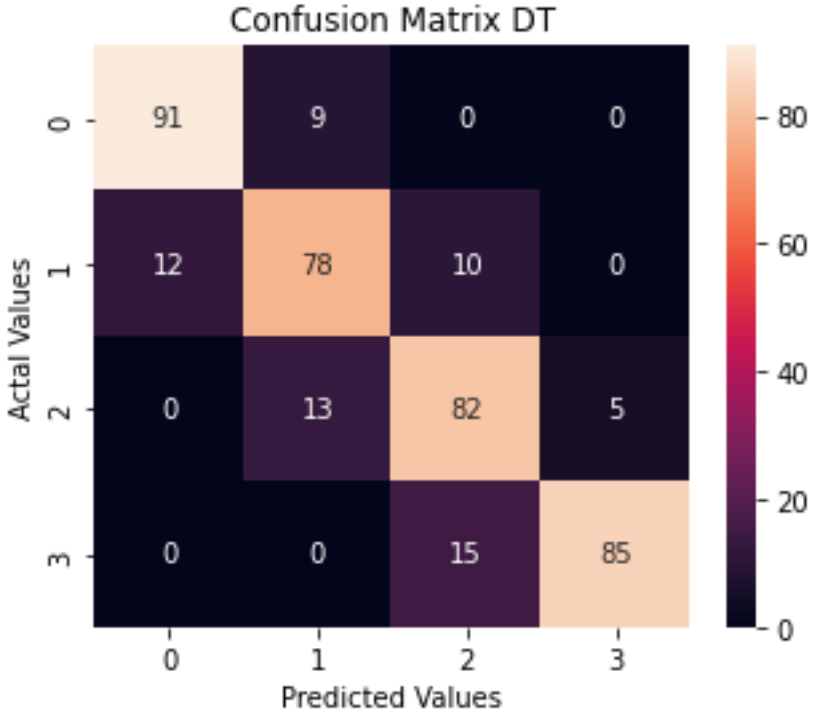
معیار criterion– جداساز splitter- ماکسیمم عمق max_depth

مدل ها از OVA استفاده کرده اند.

۵. خیر زیرا بدیهی است که ممکن است تشخیص یک کلاس سخت تر از کلاس دیگر باشد یا در داده تست نمونه هاس سخت تری از کلاس هاس متفاوت وجود داشته باشد.



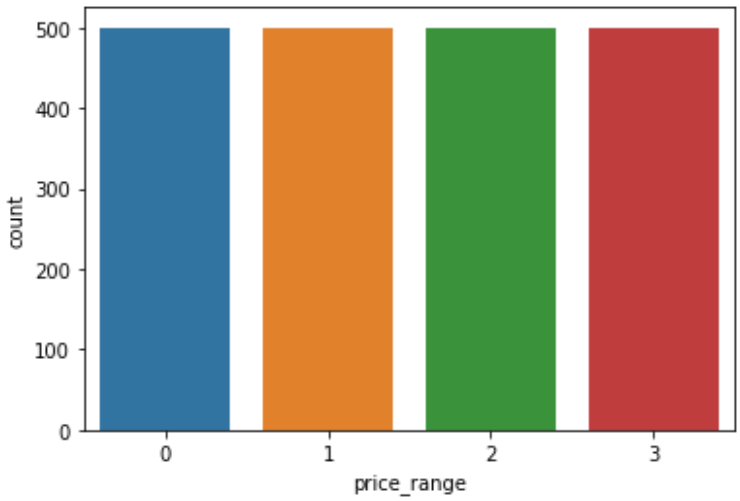

Confusion Matrix SVM
1e+02
0
0
0
2
97
1
0
0
7
89
4
0
0
3
97
Actal Values
Predicted Values
0
1
2
3
100
80
60
40
20
0


Confusion Matrix DT
91
9
0
0
12
78
10
0
0
13
82
5
0
0
15
85
Actal Values
Predicted Values
0
1
2
3
80
60
40
20
0

۶.

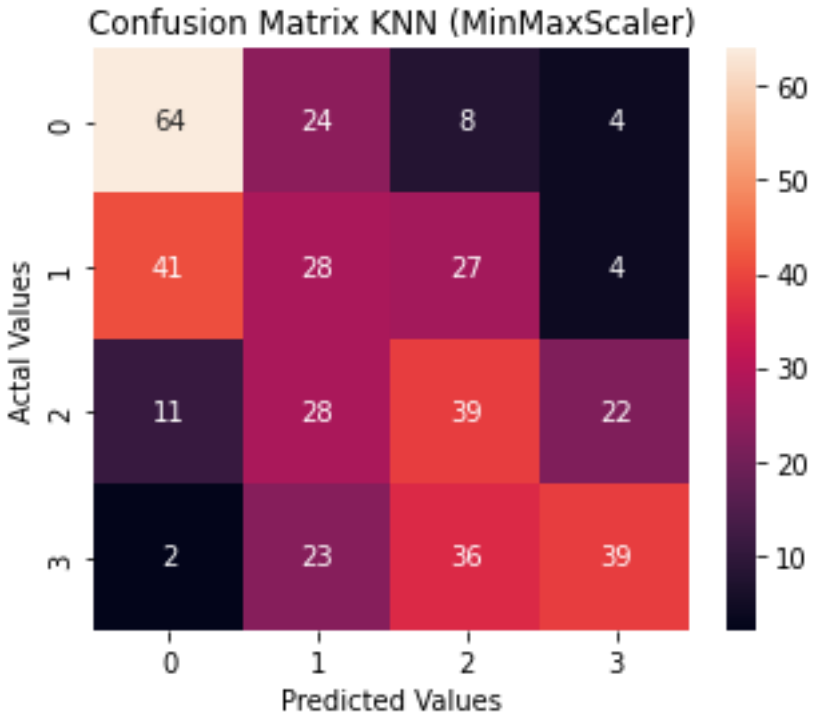
بله می بینیم که از کلاس های متفاوت مقدار تقریبا یکسانی وجود دارد و دیتاست متوازن است

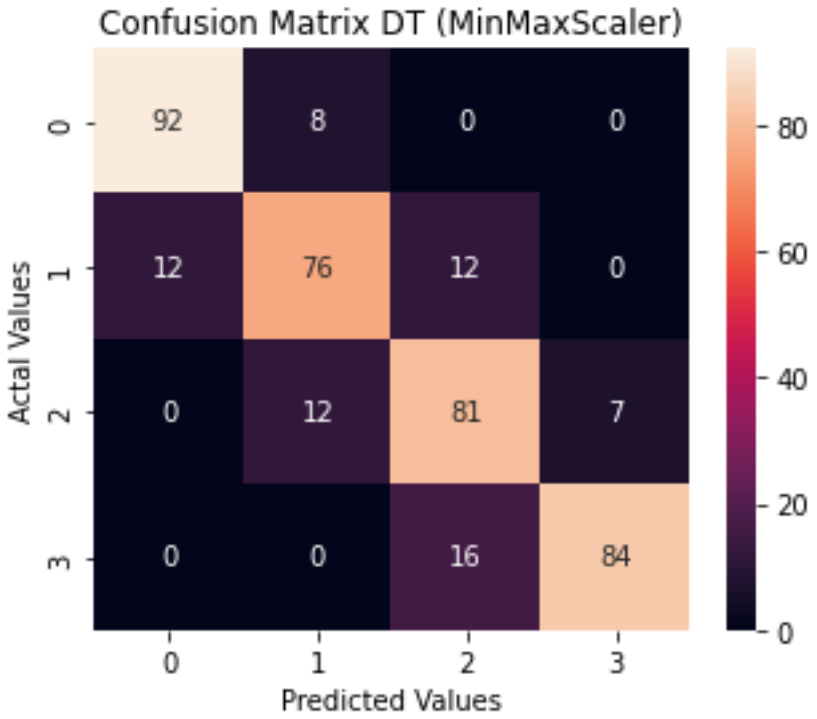
اگر دیتاست متوازن نبود میشد با روش هایی مانند روش SMOTE از کلاس کوچکتر داده هایی شبیه به آن کلاس تولید کرد.

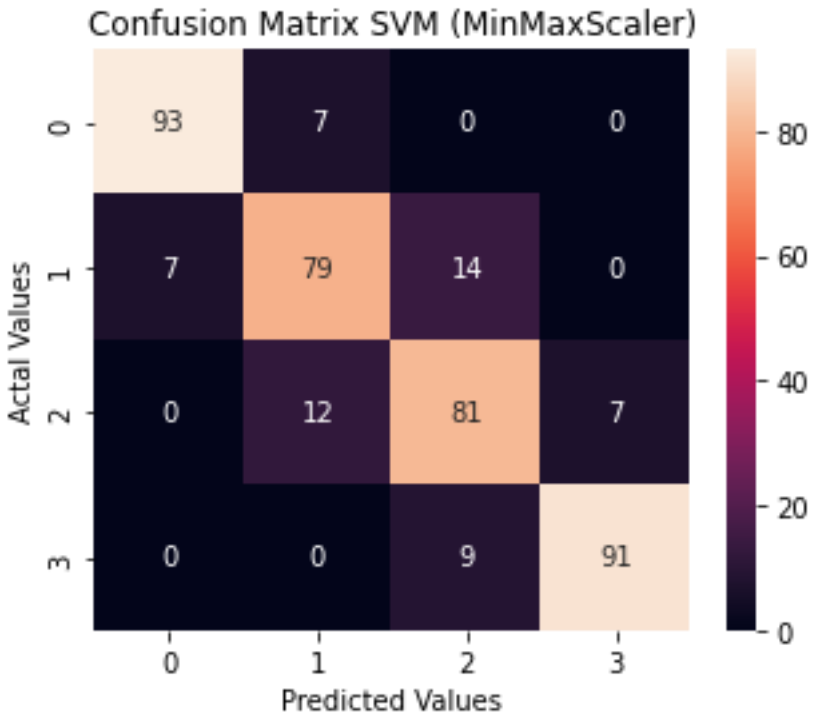


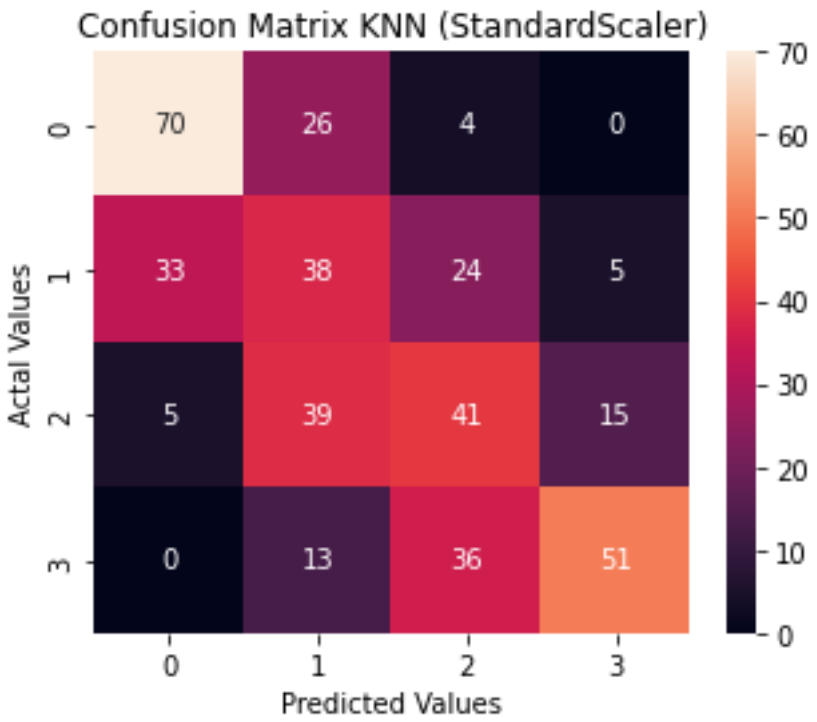
۷. با بررسی جدول در هم ریختگی دیتاست scale شده توسط دو روش در می یابیم که scaling وابسته به مدل است

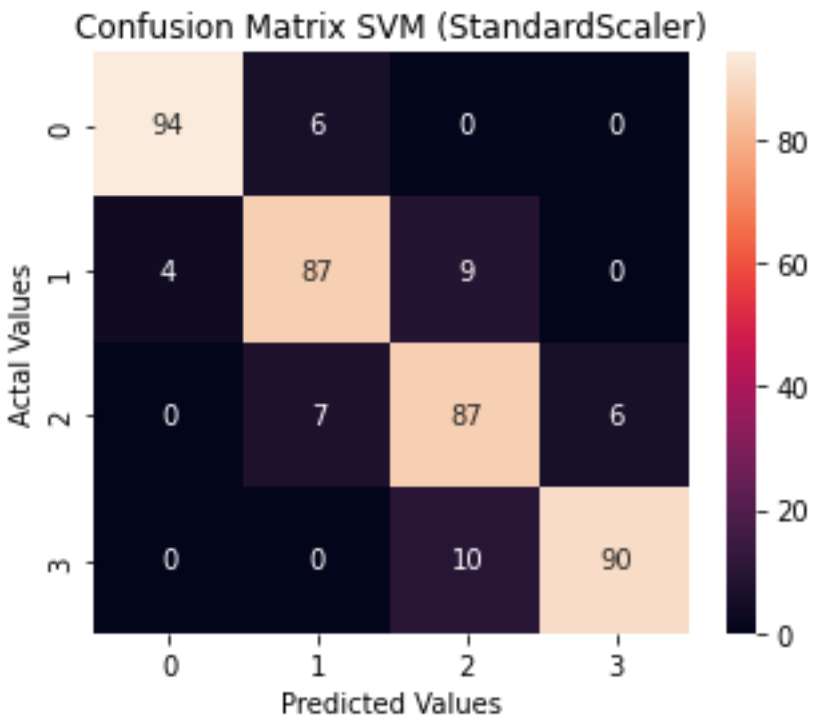
به نحوی که در knn باعث بدتر شدن نتایج نیز شده است و در سایرین تفاوت کمی ایجاد کرده است.

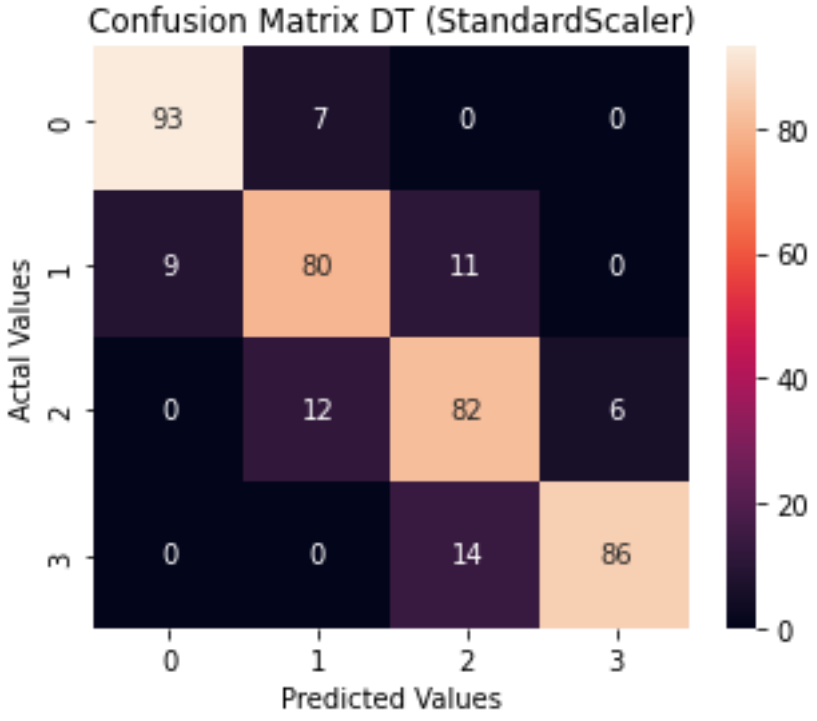





Confusion Matrix SVM (MinMaxScaler)
93
7
0
0
7
79
14
0
0
12
81
7
0
0
9
91
Actal Values
Predicted Values
0
1
2
3
80
60
40
20
0


Confusion Matrix KNN (StandardScaler)
70
26
4
0
33
38
24
5
5
39
41
15
0
13
36
51
Actal Values
Predicted Values
0
1
2
3
70
60
50
40
30
20
10
0


Confusion Matrix SVM (StandardScaler)
94
6
0
0
4
87
9
0
0
7
87
6
0
0
10
90
Actal Values
Predicted Values
0
1
2
3
80
60
40
20
0


Confusion Matrix DT (StandardScaler)
93
7
0
0
9
80
11
0
0
12
82
6
0
0
14
86
Actal Values
Predicted Values
0
1
2
3
80
60
40
20
0

-۸-۹-۱۰

نتایج سه سوال را برای راحتی مقایسه در یک جدول آوردیم.

جدول نشان می دهد که svm نسبت به سایر مدل ها عملکرد بهتری از خود نشان داده است و در pca با pov های ۰,۹۹و ۰,۹۵ که چهار ستون بدست می آید بهترین عملکرد را از خود نشان داده است

در بخش متوازن سازی نیز میبینیم که متوازن کردن باعث بهبود نتایج نسبت به زمان نامتوازنی شده است.

مدل	Accuracy	precision	recall	f1-score
KNN	0.94	0.93	0.93	0.93
SVM	**0.96**	**0.96**	**0.96**	**0.96**
DT	0.84	0.84	0.84	0.84
KNN (MinMaxScaler)	0.42	0.43	0.43	0.42
SVM (MinMaxScaler)	0.86	0.86	0.86	0.86
DT (MinMaxScaler)	0.83	0.83	0.83	0.83
KNN (StandardScaler)	0.50	0.52	0.50	0.51
SVM (StandardScaler)	0.90	0.90	0.90	0.90
DT (StandardScaler)	0.85	0.86	0.85	0.85
KNN (n_components=2)	0.80	0.80	0.80	0.80
SVM (n_components=2)	0.82	0.83	0.82	0.83
DT (n_components=2)	0.74	0.74	0.74	0.74
KNN (n_components=3)	0.94	0.94	0.94	0.94
SVM (n_components=3)	**0.96**	**0.96**	**0.96**	**0.96**
DT (n_components=3)	0.91	0.91	0.91	0.90
KNN (n_components=4)	0.93	0.93	0.93	0.93
SVM (n_components=4)	**0.96**	**0.96**	**0.97**	**0.96**
DT (n_components=4)	0.91	0.91	0.91	0.91
KNN (unbalanced)	0.96	0.96	0.96	0.96
SVM (unbalanced)	0.97	0.97	0.98	0.97
DT (unbalanced)	0.89	0.87	0.87	0.87
KNN (balanced)	0.96	0.96	0.96	0.96
SVM (balanced)	0.96	0.97	0.96	0.96
DT (balanced)	0.90	0.90	0.90	0.90

تسک های امتیازی

- از dask استفاده شده است.

دیتاست ۲

۱.

دیتاست دارای مقادیر زیاد از دست رفته به شرح زیر است .(به درصد)

```
telekomHybridUploadSpeed    0.832546
electricityKwhPrice         0.825754
electricityBasePrice        0.825754
energyEfficiencyClass       0.710668
lastRefurbish               0.699792
heatingCosts                0.681912
noParkSpaces                0.653889
petsAllowed                 0.426160
interiorQual                0.419063
thermalChar                 0.396154
numberOfFloors              0.363519
houseNumber                 0.264155
streetPlain                 0.264136
condition                   0.254748
yearConstructed             0.212182
yearConstructedRange        0.212182
firingTypes                 0.211880
facilities                  0.196853
floor                       0.190846
heatingType                 0.166844
totalRent                   0.150705
typeOfFlat                  0.136187
telekomUploadSpeed          0.124077
telekomTvOffer              0.121328
description                 0.073450
serviceCharge               0.025698
pricetrend                  0.006814
regio3                      0.000000
regio2                      0.000000
livingSpaceRange            0.000000
garden                      0.000000
noRoomsRange                0.000000
regio1                      0.000000
noRooms                     0.000000
geo_plz                     0.000000
baseRentRange               0.000000
lift                        0.000000
street                      0.000000
geo_krs                     0.000000
livingSpace                 0.000000
baseRent                    0.000000
cellar                      0.000000
geo_bln                     0.000000
hasKitchen                  0.000000
scoutId                     0.000000
picturecount                0.000000
balcony                     0.000000
newlyConst                  0.000000
date                        0.000000
dtype: float64
```

- همانطور که می بینیم بعضی از ویژگی ها ۸۰ درصدشان از دست رفته هست و ترمیم این ها ممکن است مدل را دچار سوگیری کند پس ویژگی هایی که بیشتر از نصف آن ها دارای مقادیر از دست رفته بود را حذف کردیم.

- در گام بعدی داده های تکراری را بررسی کردیم که خوشبختانه داده تکراری موجود نبود.

- در گام بعدی آن هایی که برچسب آن ها دارای مقادیر صفر یا خالی بود را حذف کردیم چون تارگت تاثیر بیشتری بر روی مدل را داراست.

- در گام بعدی خانه هایی که متراژ آن ها صفر بود را نیز حذف کردیم که این ها داده اشتباه به حساب می آیند.

- در گام بعد ویژگی هایی مانند تاریخ آگهی، آی دی ، شماره پلاک خانه که تاثیری در قیمت ندارند را حذف کردیم.

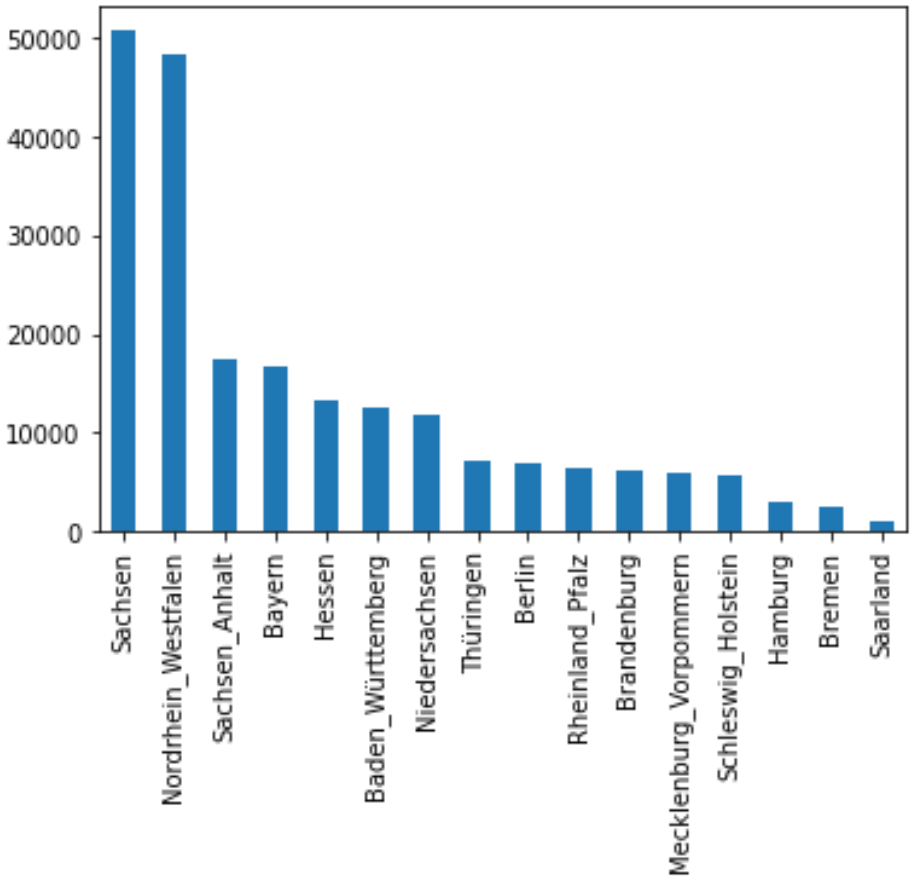
- در گام بعد داده های پرت را با روش IQR به روش multi processing حذف کردیم.

- در گام بعد ویژگی های غیر عددی را با مد به روش multi processing پر کردیم.

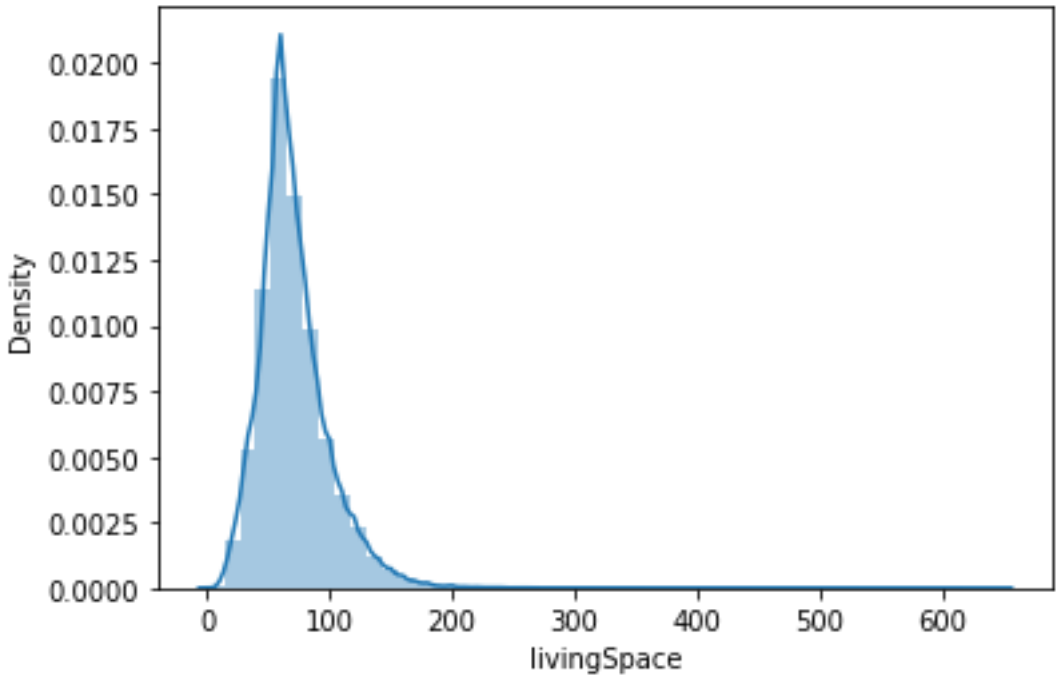
- داده های غیر عددی که بیش از ۱۰۰ نوع داشتند را حذف کردیم. (مانند کوچه)

در این مرحله دیگر هیچ داده خالی نداریم.

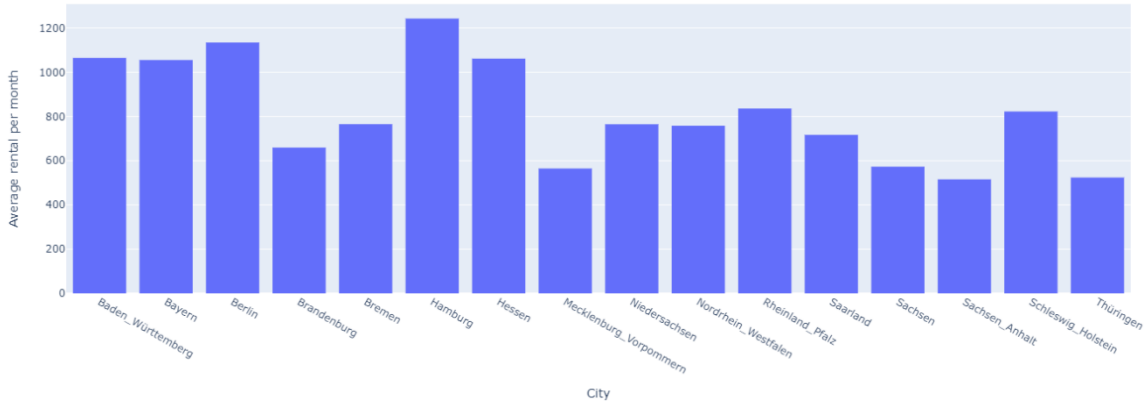
۲. بر حسب مناطق تعداد خانه ها را بدست آوردیم که در Sachsen بیشترین خانه وجود دارد.



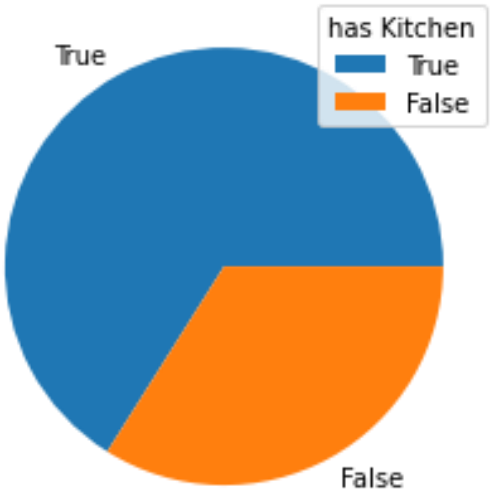
نمودار زیر نشان می دهد که بیشتر خانه ها حدود ۷۰ متر دارند



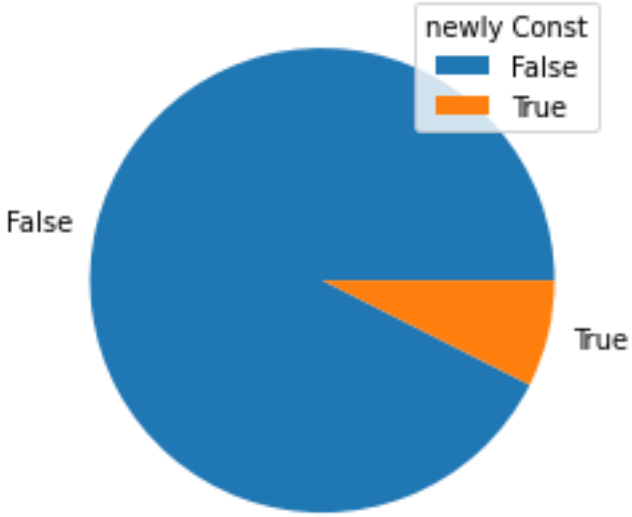
نمودار زیر نیز قیمت ماینگین اجاره در هر منطقه را نشان می دهد.



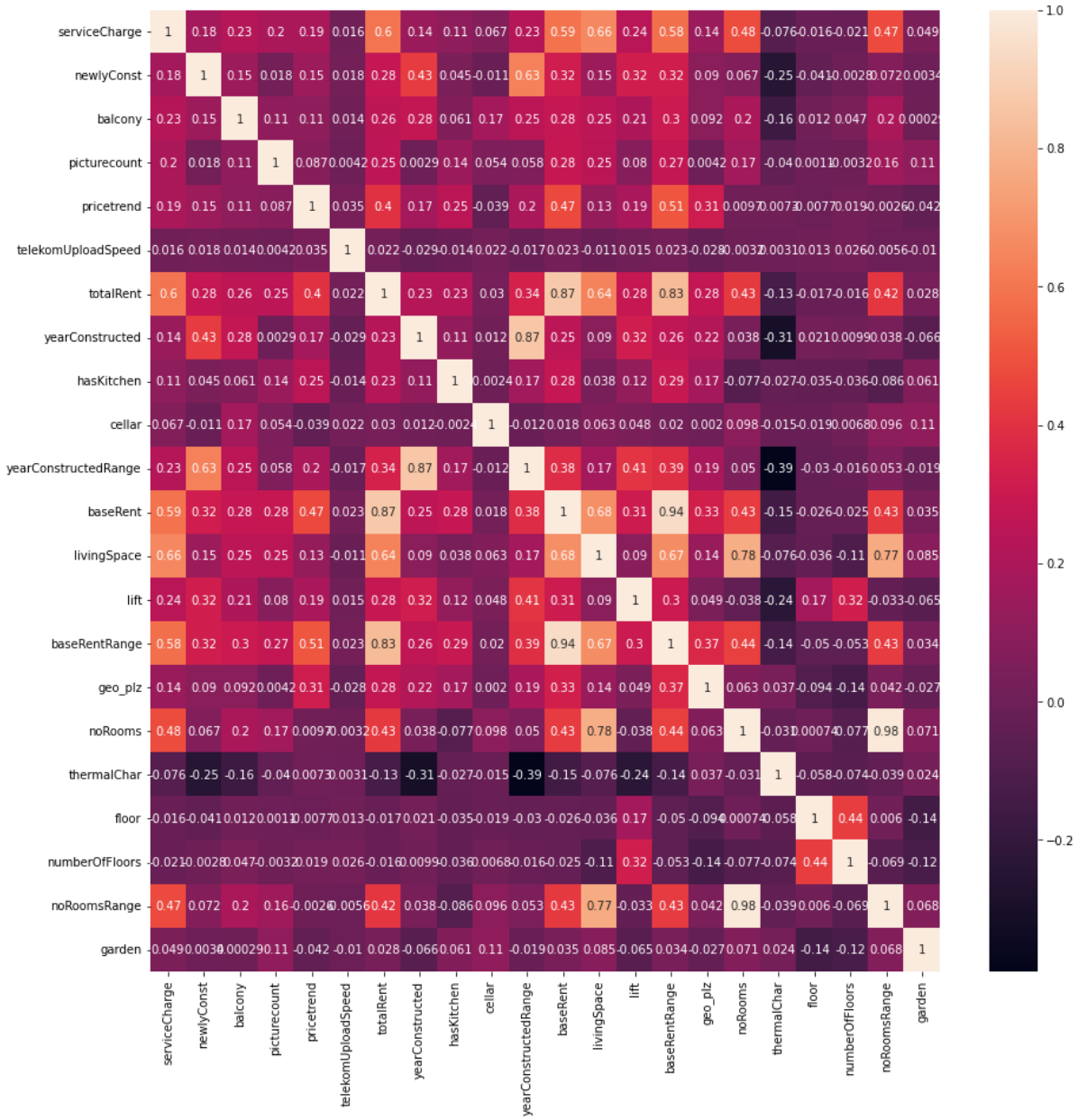
نمودار زیر نیز تعداد خانه هایی که آشپزخانه دارندرا نسبت به خانه های بدون آشپزخانه نشان می دهد.



نمودار زیر نیز تعداد خانه هایی که تازه ساخت هستند را نسبت به خانه های قدیم ساخت نشان می دهد.



نمودار زیر نیط ارتباط ویژگی ها را با یکدیگر نشان می دهد.



.۳

ابتدا ویژگی های غیر عددی را تبدیل کرده و سپس به نسبت ۸۰ به ۲۰ داده آموزش و تست جدا کرده و با رگرسیون خطی به مدلسازی پرداختیم

ارزیابی مدل با خطای MAE برابر با ۲۷,۳۹۲۶۵۲۵۲۸۰۲۳۲۲۷ است .

۴و۵ . چون این دو سوال به بخش پیش پردازش مربوط است در آن بخش انجام شده است.

تسک های امتیازی

با روش انتخاب ویژگی بازگشتی به انتخاب ویژگی پرداختیم که این روش حدودا نصف ویژگی ها را ارزشمند تلقی کرد.

ارزیابی مدل با خطای MAE روی دیتاست انتخاب ویزگی شده برابر با ۹۷,۸۴۴۰۹۲۱۴۰۹۹۳۰۳ است .

خطا به مقدار قابل قبولی افزایش یافته است اما باید در نظر داشت که که این نتیجه روی نصف ویژگی بدست آمده و زمان آموزش و ارزیابی را به نصف کا هش داده است.