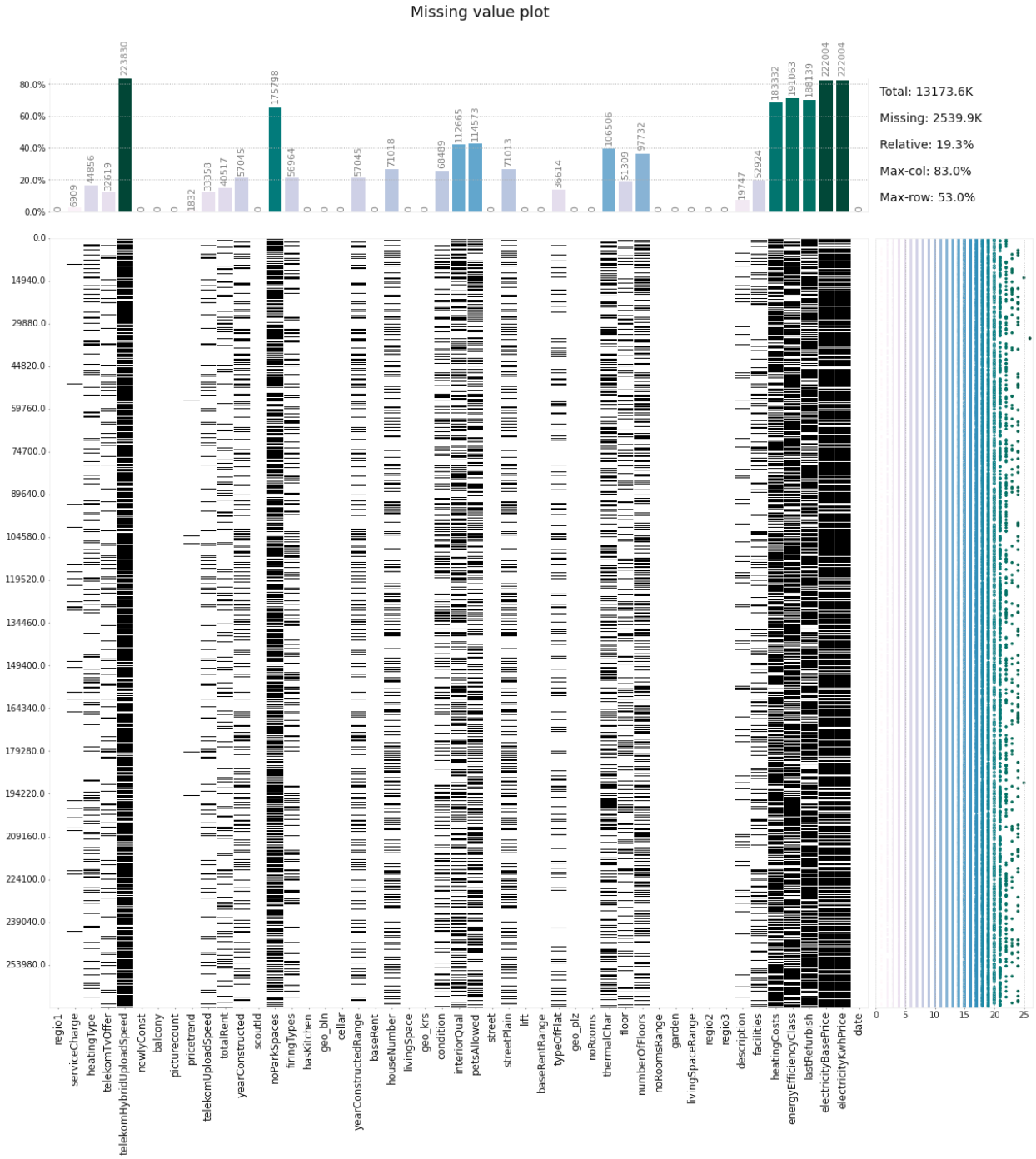
داده های این تمرین مربوط به خانه های آلمان می شود.

پاک سازی داده ها

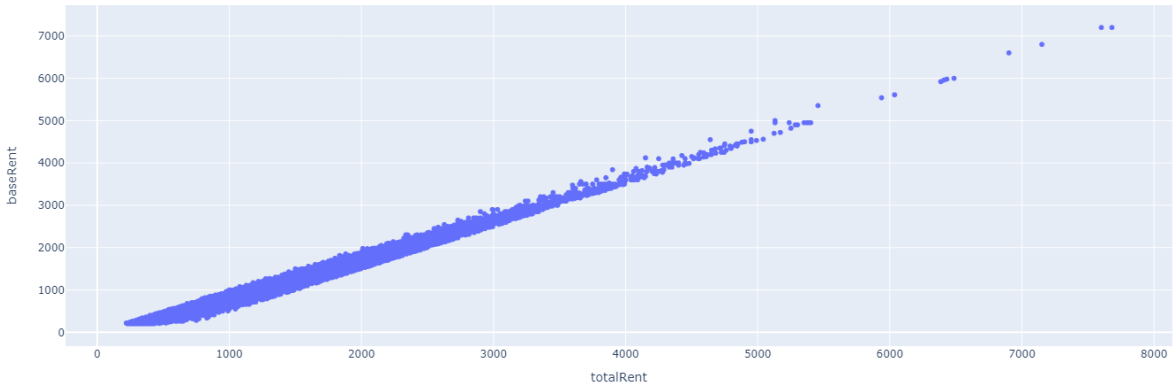


همان طور که در شکل بالا مشاهده می شود دیتا ما دارای قسمت های خالی زیادی هست. ئس نیاز به پاک سازی داریم. مراحل کار به صورت زیر است.

1. همه ویژگی هایی که 20 درصد به بالا missing value دارند را حذف می کنیم.
2. ویژگی های که به نظر نیاز نداریم را حذف می کنیم. این ویژگی ها به نظر نسبت به بقیه باید تاثیر کم تری داشته باشند. لیست ویژگی ها:

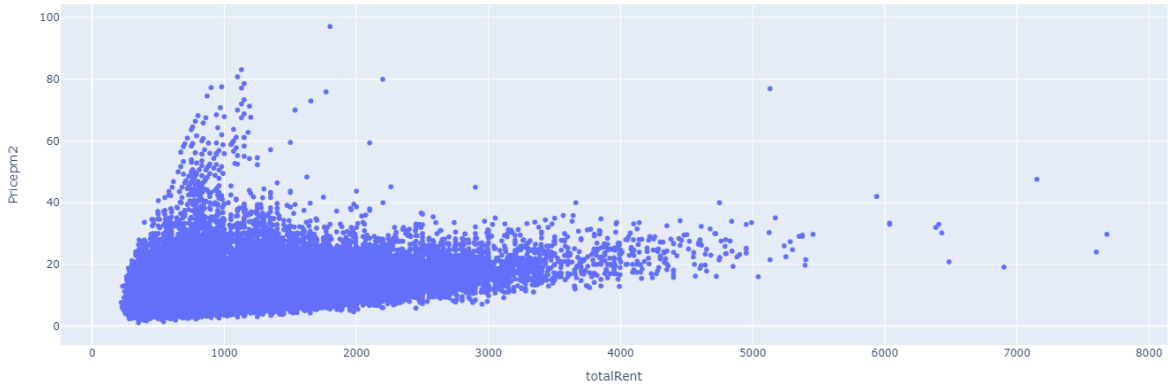
```
['livingSpaceRange', 'street', 'description', 'facilities', 'geo_krs',
'geo_plz', 'scoutId', 'regio1', 'telekomTvOffer', 'pricetrend', 'regio3',
'noRoomsRange', 'picturecount', 'geo_bln', 'date', 'houseNumber',
'streetPlain', 'firingTypes', 'yearConstructedRange']
```

3. ستون condition. در مقدار های خالی این ستون مقدار other قرار می دهیم.
4. ستون year constructed را هم با میانگین داده های با condition یکسان پر می کنیم.
5. ستون جدیدی اضافه می کنیم که زمان اخرین بازسازی یا ساخت خانه را نشان می دهد. یعنی این که چه مقدار از آن تاریخ گذشته است.
6. 20 شهر با بیشترین تعداد داده را بر می داریم و بقیه داده ها را حذف می کنیم. با این کار مدل خود را فقط برای همین 20 شهر آموزش خواهیم داد. این شهر ها که حذف شدند کم تر از 2000 داده دارند.
7. حذف outlier ها. به نمودار های زیر توجه کنید.



همان طور که مشاهده می کنید به نظر بیشتر از 5500 داده ها outlier هستند.

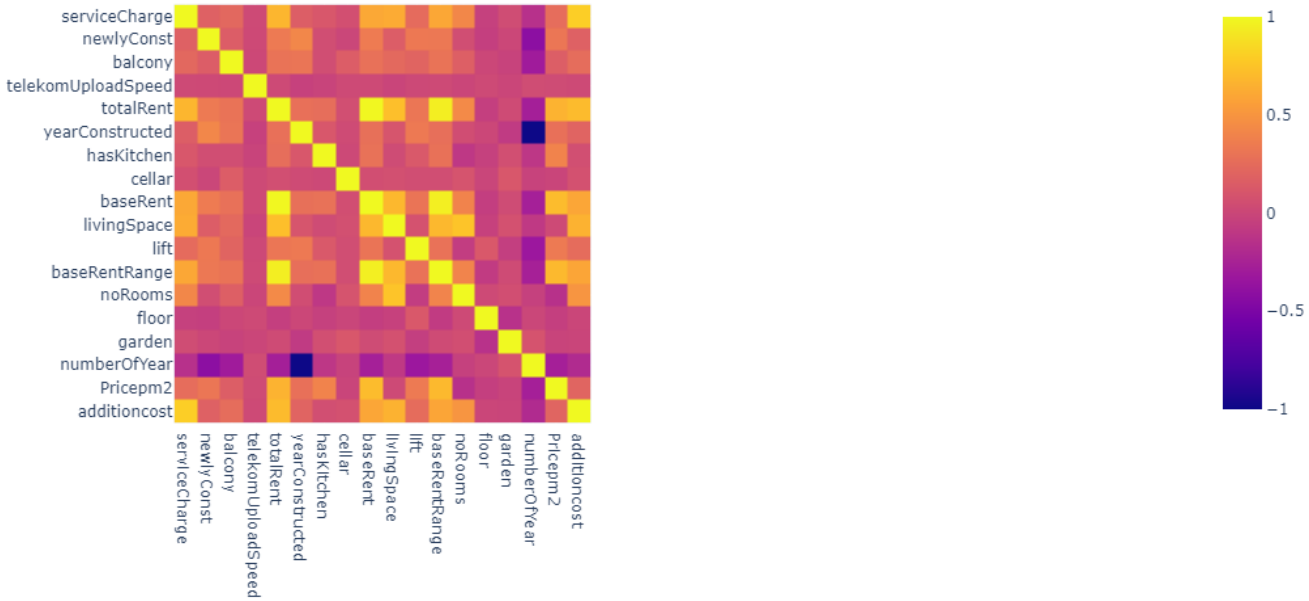
8. یک ستون جدید درست می کنیم که قیمت براساس هر متر را نشان می دهد. نمودار مربوط با آن این شکل می شود.



9. ستون های heating type و type of flat را که مقدار ندارند با استفاده از مد داده ها پر می کنیم.

10. ستون base rent را با توجه به corr بالا با total rent حذف می کنیم.

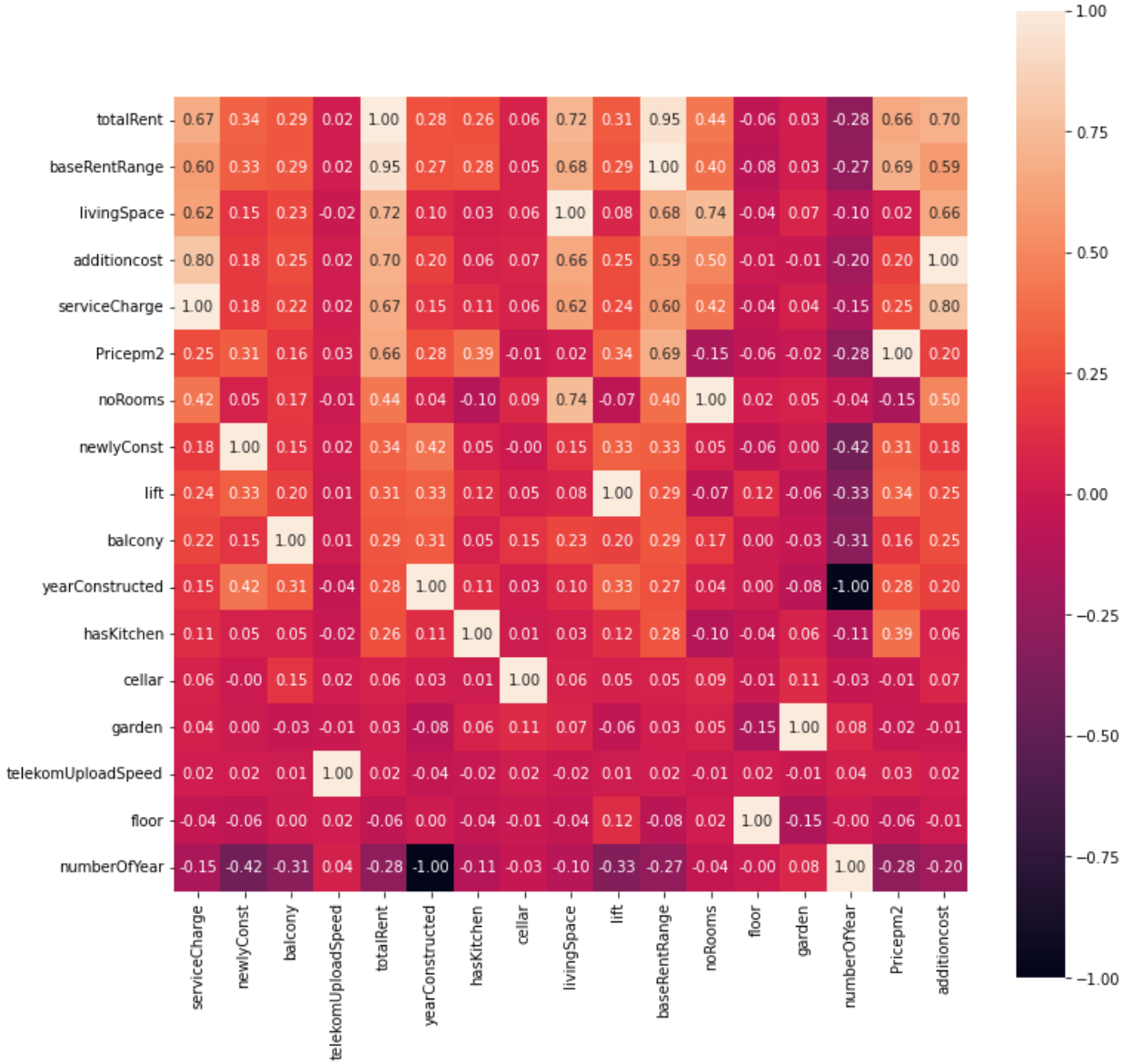
تا این جا نمودار correlation داده ها به این صورت می شود.



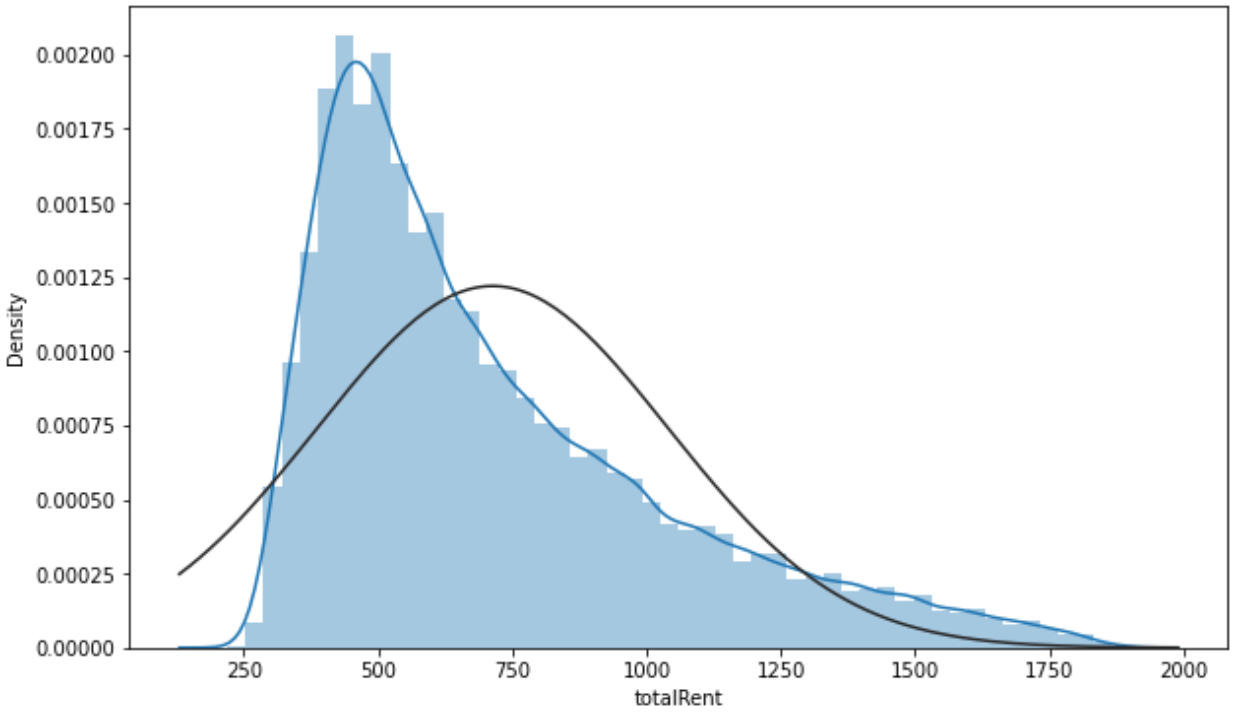
11. فقط ستون telecomUploadSpeed یک سری داده خالی دارد که آن ها را پاک می کنیم.

مصور سازی داده ها

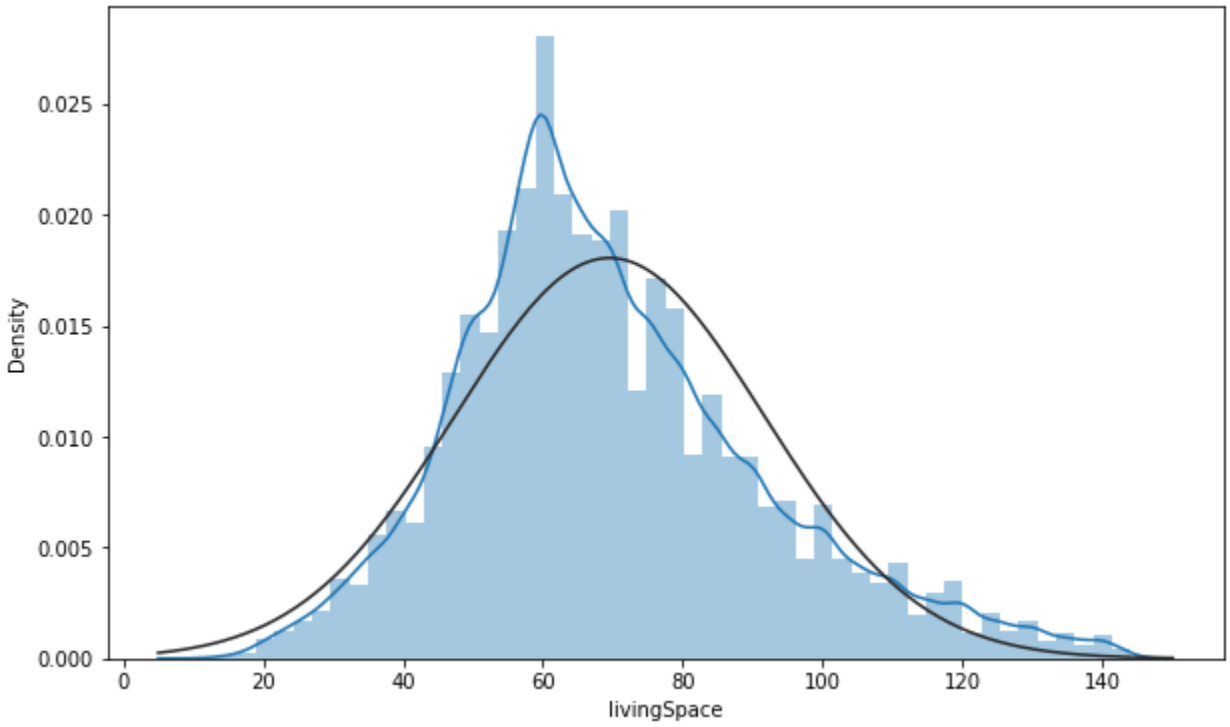
نمودار corr به این صورت می شود.



حال نمودار total rent distribution را می کشیم.



نمودار زیر نمودار distribution برای متغیر living space است.

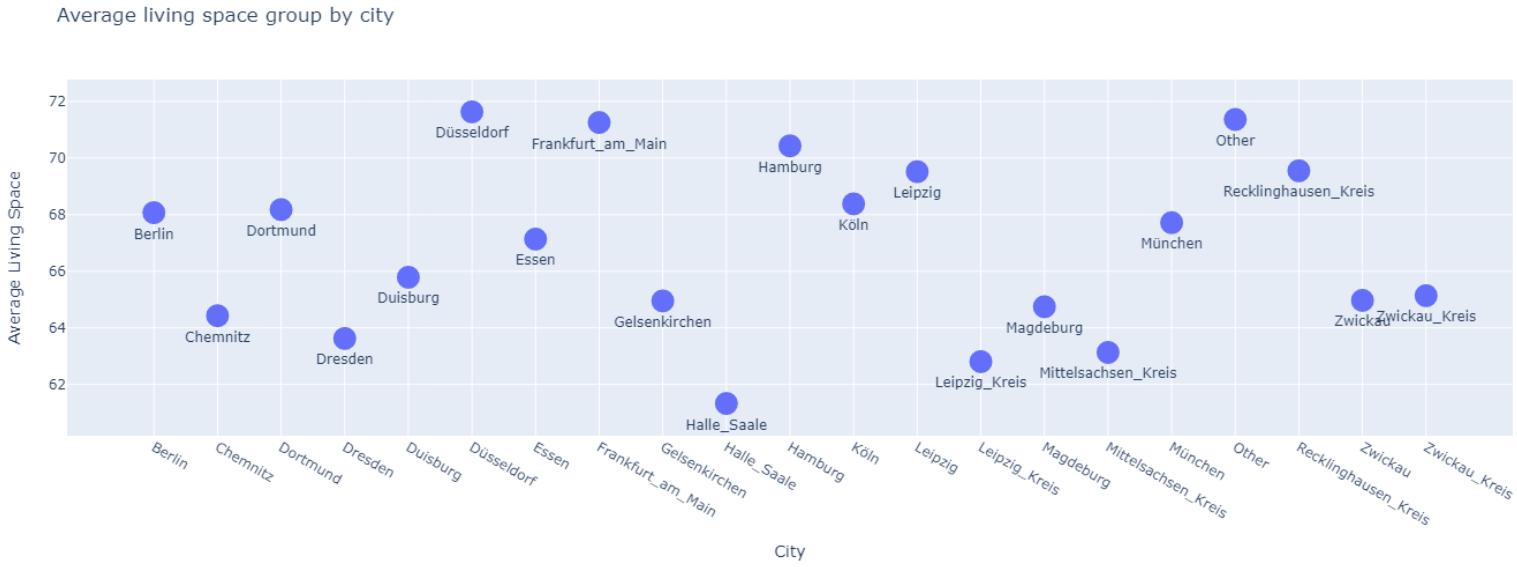


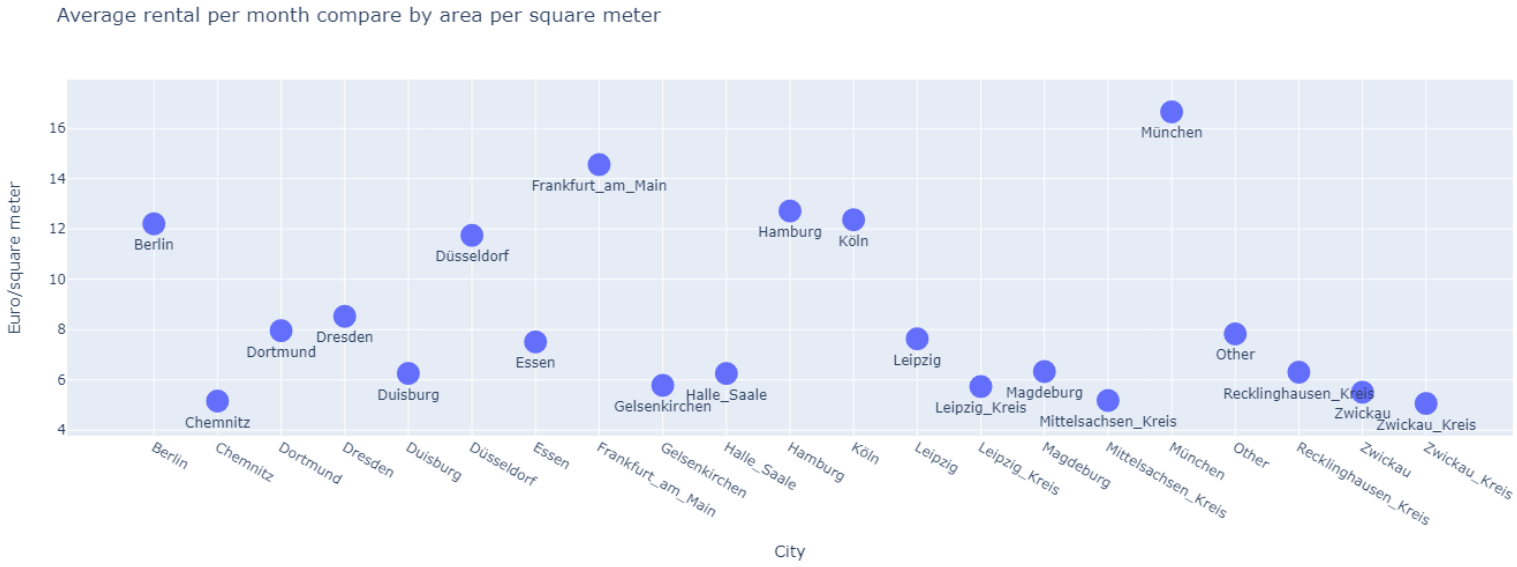
نمودار بعدی نمودار شهر ها است.

Pie chart of all the City ratio in the dataset

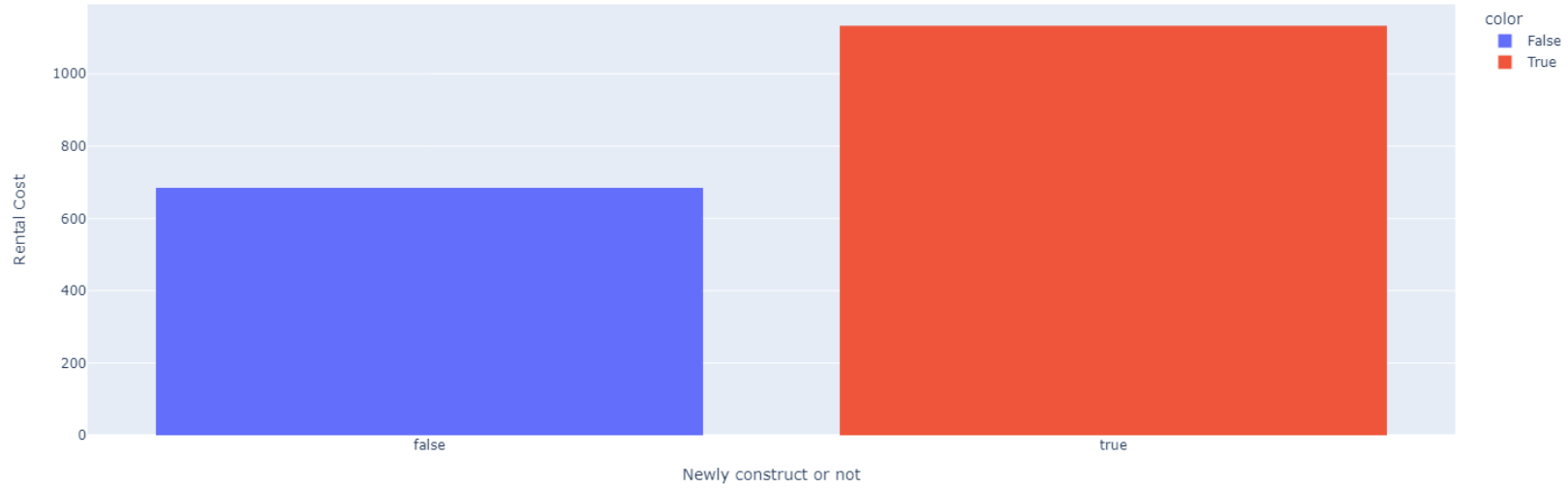
Pie chart of all the City ratio in the dataset exclude 'Other'

نمودار های زیر هم نمودار های مختلف بر اساس شهر ها است.





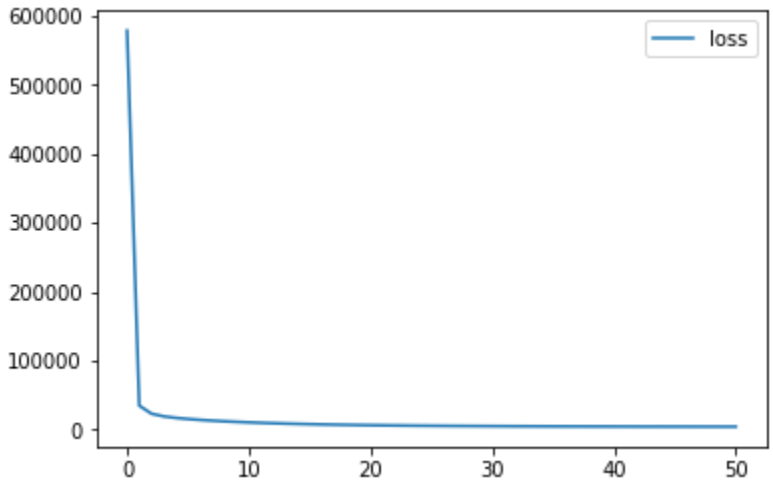
نمودار قیمت بر اساس این که خانه تازه است یا خیر



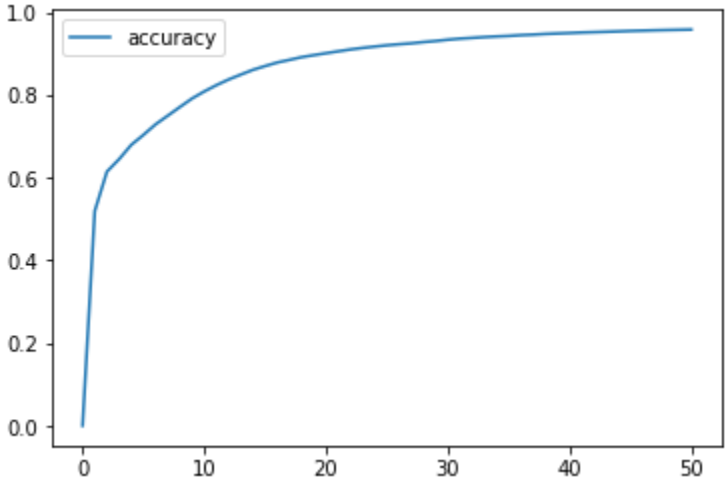
ساخت مدل

مدل با توجه به سوال باید با 3 ویژگی ساخته شود. صرفا ویژگی heatingType را onehot encode کردیم.

مدل رگرسیون خطی را پیاده سازی کردیم و خروجی خطا ها در هر epoch به صورت زیر است.



نمودار دقت مدل هم به شکل زیر است.



در انتها مدل پیاده سازی شده را با مدل های آماده و ridge و lasso مقایسه می کنیم.

model	accuracy
Linear reg Imp	0.9579643505903724
Linear reg package	0.9629412867037032
Ridge	0.9629412836303942
Lasso	0.961634934057433

علت اختلاف مدل ئیاده سازی شده و همان مدل با package میتواند در تعداد iteration ها باشد. این مقدار در مدل ما 50 بود. و اما در کل اختلاف انواع مدل ها به خاطر تفاوت در محاسبه آن ها است.