**سوالات اصلی پروژه:**

1- قیمت بیشترین و کمترین آن چقدر است؟

2- عوامل موثر بر قیمت کدام هستند؟

3- کدام عوامل میتوان خارج کرد؟

**سوال یک:**

قیمت بیشترین و کمترین آن چقدر است؟

**count 2.161300e+04**

**mean 5.400881e+05**

**std 3.671272e+05**

**min 7.500000e+04**

**25% 3.219500e+05**

**50% 4.500000e+05**

**75% 6.450000e+05**

**max 7.700000e+06**

**Name: price, dtype: float64**

**سوال دوم:**

عوامل موثر بر قیمت کدام هستند؟

با توجه به اینکه df.corr نسبت به قیمت بگیریم 10 تای اول آن به این شکل است

['price', 'sqft\_living', 'grade', 'sqft\_above', 'sqft\_living15',

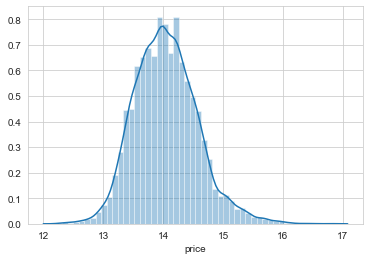
'bathrooms', 'view', 'sqft\_basement', 'bedrooms', 'lat'],

dtype='object')

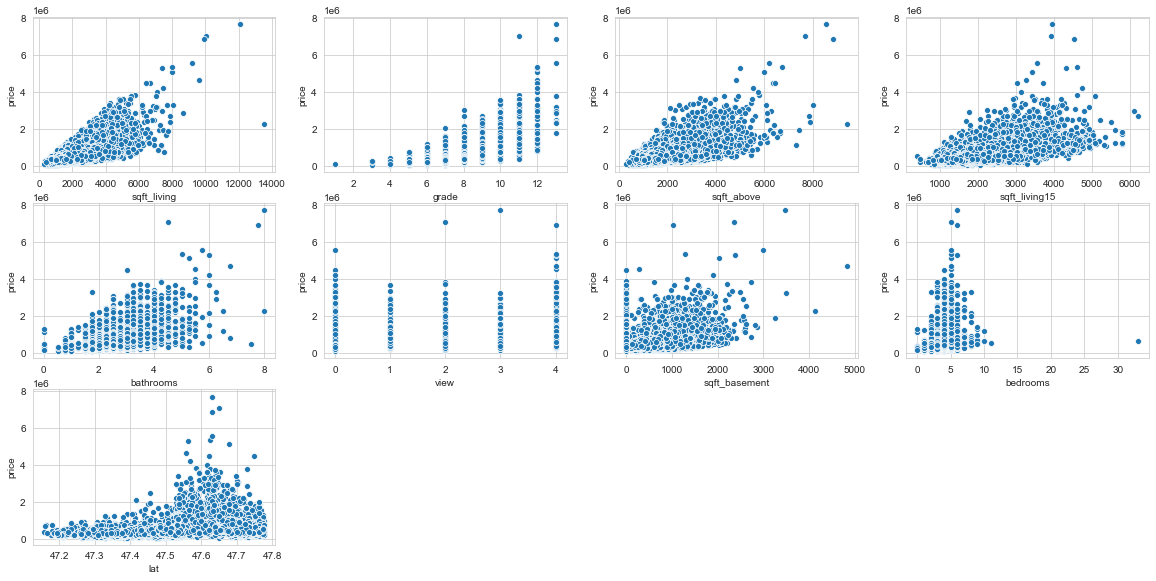
**سوال سوم:**

پخش شدگی قیمت به چه صورت است؟

همانطور که می بینی قیمت تا 1000 بیشترین تعداد را دارد که اگر لگاریتم آن را بگیریم روی 14 می باشد



حالا اگر این داده های مرتبط را با هم بسنجیم همچین نتیجه ای بدست می آید



پخش شدگی داده ها در

['price', 'sqft\_living', 'grade', 'sqft\_above', 'sqft\_living15',

'bathrooms', 'sqft\_basement', 'bedrooms', 'lat'], میبینیم ولی نکته جالب در میزان اتاق ها هست.

**سوال چهارم:**

بیشترین سهم اتاق ها چقدر است؟

بیشترین سهم اتاق ها 3 تایی و 4تایی و 2 تایی است که به شرح زیر است

3 45.454125

4 31.841947

2 12.770092

5 7.407579

6 1.258502

1 0.920742

7 0.175820

0 0.060149

8 0.060149

9 0.027761

10 0.013881

11 0.004627

33 0.004627

**سوال پنجم:**

بیشترین سهم فیچرهای عددی به چه صورت است؟

تمام فیچرهای عددی را جدا میکنیم که به شرح زیر است

counts\_num = ['bedrooms', 'bathrooms','floors','view','condition', 'grade']

که value\_count آن به شرح زیر است

the value counts of bedrooms is

3 45.454125

4 31.841947

2 12.770092

5 7.407579

6 1.258502

1 0.920742

7 0.175820

0 0.060149

8 0.060149

9 0.027761

10 0.013881

11 0.004627

33 0.004627

Name: bedrooms, dtype: float64

the value counts of bathrooms is

2.50 24.892426

1.00 17.822607

1.75 14.102623

2.25 9.471152

2.00 8.929811

1.50 6.690418

2.75 5.482811

3.00 3.484014

3.50 3.382224

3.25 2.725212

3.75 0.717161

4.00 0.629251

4.50 0.462684

4.25 0.365521

0.75 0.333133

4.75 0.106417

5.00 0.097164

5.25 0.060149

0.00 0.046268

5.50 0.046268

1.25 0.041642

6.00 0.027761

0.50 0.018507

5.75 0.018507

6.75 0.009254

8.00 0.009254

6.25 0.009254

6.50 0.009254

7.50 0.004627

7.75 0.004627

Name: bathrooms, dtype: float64

the value counts of floors is

1.0 49.414704

2.0 38.129829

1.5 8.837274

3.0 2.836256

2.5 0.744922

3.5 0.037015

Name: floors, dtype: float64

the value counts of view is

0 90.172581

2 4.455652

3 2.359691

1 1.536113

4 1.475964

Name: view, dtype: float64

the value counts of condition is

3 64.919262

4 26.275852

5 7.870263

2 0.795817

1 0.138805

Name: condition, dtype: float64

the value counts of grade is

7 41.553695

8 28.075695

9 12.099200

6 9.429510

10 5.246842

11 1.846111

5 1.119696

12 0.416416

4 0.134179

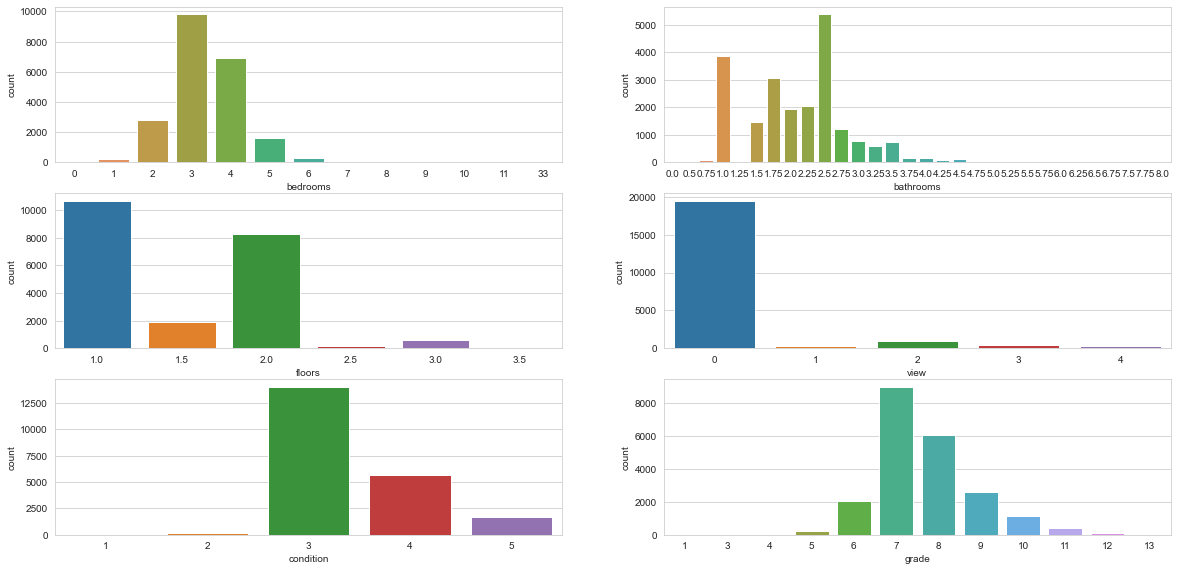
13 0.060149

3 0.013881

1 0.004627

Name: grade, dtype: float64

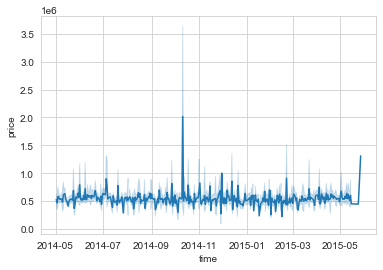
که روی نمودار هم برود به شکل زیر است



**سوال ششم:**

تاثیر زمان به چه صورت است؟

تو مقطعی از سال 2014 افزایش قیمت داشته ایم که عوامل دیگر داخل آن تاثیرگذار میتواند باشد.



**سوال هفتم:**

میانگین زمان به چه صورت است در سال؟

در سال 2014 و 2015 میانگین قیمت به شرح زیر است.

year

2014 539181.428415

2015 541988.992264

Name: price, dtype: float64

میانگین زمان در طی ماه به صورت زیر است:

1 525870.889571

2 507851.371200

3 543977.187200

4 561837.774989

5 550768.785833

6 558002.199541

7 544788.764360

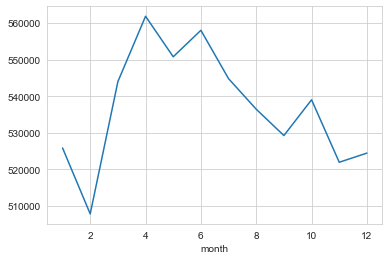
8 536445.276804

9 529253.821871

10 539026.971778

11 521961.009213

12 524461.866757

****

ماه چهارم افزایش قیمت داریم یعنی فروردین گرون ترین حالت است.

**سوال هشتم:**

نال و داده های پرت چگونه است؟

با بررسی های انجام شده دادههای نال و پرتی وجود نداشت

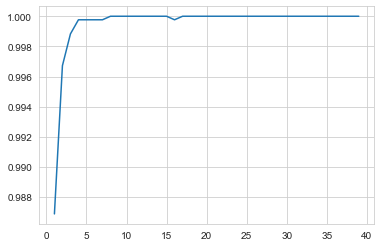
(21613, 19)

(21613, 19)

**سوال نهم:**

الگوریتم های ماشین کدام موثر بوده است؟

با استفاده از knn ابتدا فیچرهای زمان و id را حذف کردم و تمامی آن ها را استاندارد کردم که بهترین k عدد 5 بوده است.



که خطا های آن به شرح زیر است

91001.37656982192

29474452033.98911

171681.251259388

از بالا به پایین mas,mse,rmse

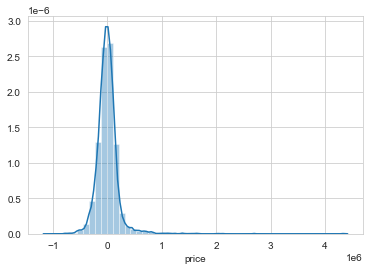
اما با رگرسیون هم آن را تست کردیم که بهترین n\_jobs شماره یک بود که خطاهای ان به شرح زیر است ( به شرطی که تابع نرمال باشد)

127682.31490032641

42936155096.290054

207210.41261551034

نکته عجیب تفاوت 3 عدد روش رگرسیون با knn بود اما اختلاف میزان مشاهده شده با پیش بینی در رگرسیون نرمال بود ه است.



حالا اگر polynomial feature را به کار بگیریم عرض از مبدا ان

15235876475571.049 میشود و میزان اسکور داده تمرین و تست به صورت زیر میشود

تابع تمرین:

0.8335663284114918

داده تست:

0.8046867010113582

و خطاهای آن به ترتیب

101860.88697210133

25051319077.589043

158276.08498313648

می شود که مدل را خیلی بهینه میکند و اگر از ridge استفاده کنیم آن گاه کمینه خطاها

101897.53019481605

25075451313.726646

158352.30125807028

و میزان واریانس اسکور ان هم 80 درصد است . اگر loasso را به کار بگیریم

101950.20103775081

25108051652.4293

158455.20392978357 هم میزانی افزایش یافت.

شبکه های عصبی با keras:

وقتی از کراس استفاده میکنیم نتایج خیلی جالب تر میشود به طوری که ما 5 لایه با 19 نرون می سازیم که activation آن relu هست و optimizer هم adam که میزان اسکور واریانس به ترتیب زیر است

98052.73694035281

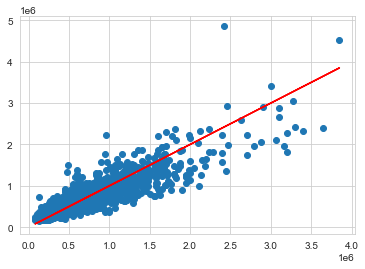
26139170875.558483

161676.1295787306

metrics.explained\_variance\_score(y\_test,pred\_i\_N)

که جواب آن

0.7966489672409514 این است



شبکه عصبی خیلی موثرتر چون lasso ,ridge بدون کمک polynomial جواب مناسبی ندارد