خوشه بندی از طریق الگوریتم  
K-Means

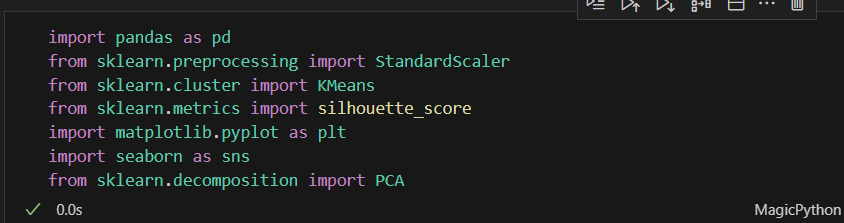
برای درس داده کاوی

# مقدمه

در این قسمت از پروژه ، توسط الگوریتم k-means خوشه های داخل داده های پاکسازی را پیدا میکنیم و شکل گرفتن ، ارتباط و مفهوم آنها را بررسی میکنیم و نمایش میدهیم.

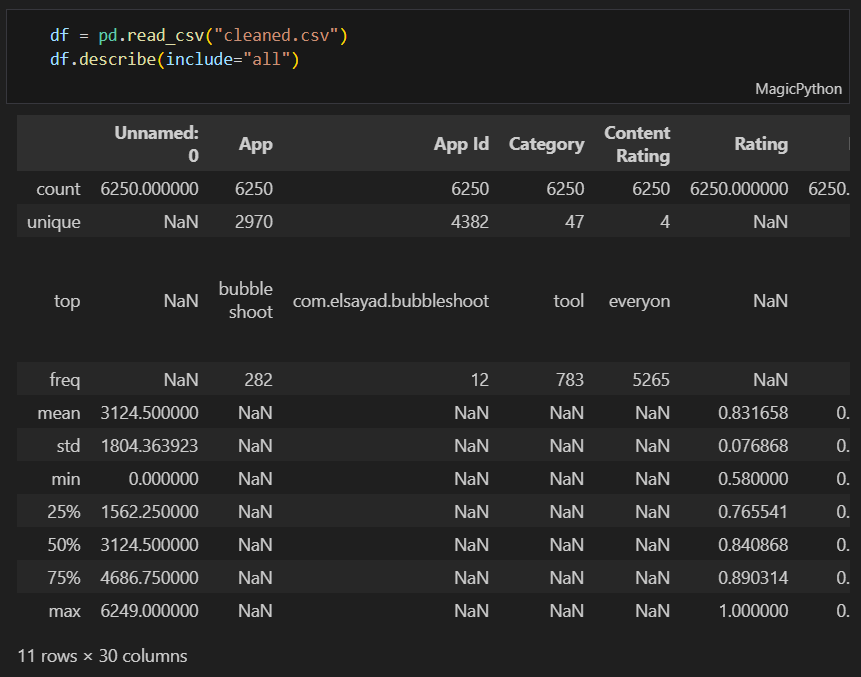
# توضیحات کد

## کتابخانه های مورد نیاز



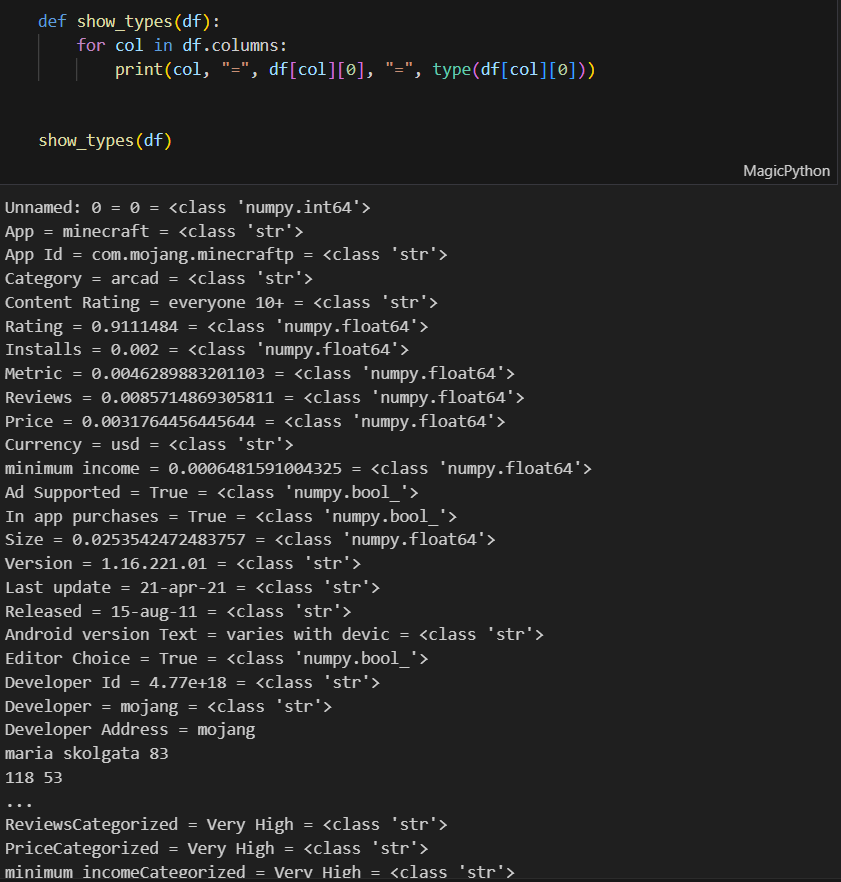
از کتابخانه pandas برای دستکاری داده ها و از کتابخانه sklearn استفاده میکینم. sklearn برای دسترسی به الگوریتم k-means که برای خوشه بندی از ان استفاده میکینم ،و همینطور دسترسی به StandardScaler که برای نرملایز کردن برای پیاده سازی الگوریتم k-means نیاز است و همینطور تابع silhouette\_score که برای بررسی کیفیت خوشه بندی است نیاز است. همینطور matplotlib و seaborn برای نمایش خوشه بندی ها استفاده می شوند.

## بارگزاری داده ها در حافظه

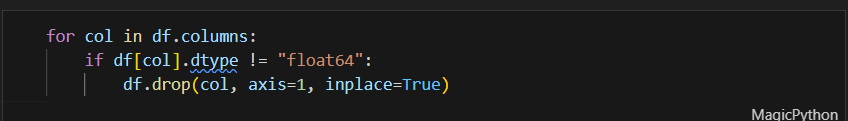
 با استفاده از کتابخانه pandas داده های پاکسازی شده در قسمت قبلی پروژه را در حافظه بارگذاری میکنیم. بعد از آن توضیحات کلی در مورد داده ها نمایش داده شده اند.

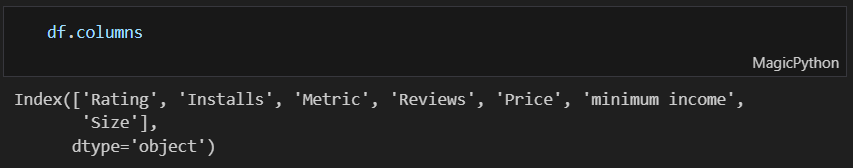
## انتخاب صفات مورد بررسی

الگوریتم k-means بر اساس فاصله میان داده ها خوشه بندی انجام میدهد و به شکل کروی . در نتیجه باید داده های عددی را بررسی کنیم.



در این قسمت تمامی انواع ستون ها ذکر شده برای بررسی.

  
در این قسمت تمامی ستون ها با مقادیر غیر عددی از حافظه حذف شده اند.



در این قسمت ستون های باقی مانده که عددی هستند و برای پیاده کردن خوشه ها استفاده حواهند شد نمایش داده شده اند.

## تبدیل به مقیاس مناسب

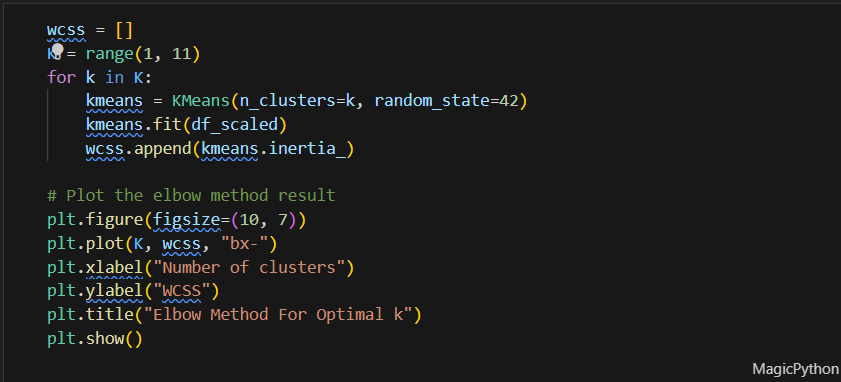
برای پیدا کردن فاصله مرکز خوشه ها، باید داده ها مقایس یکسان داشته باشند.

## 

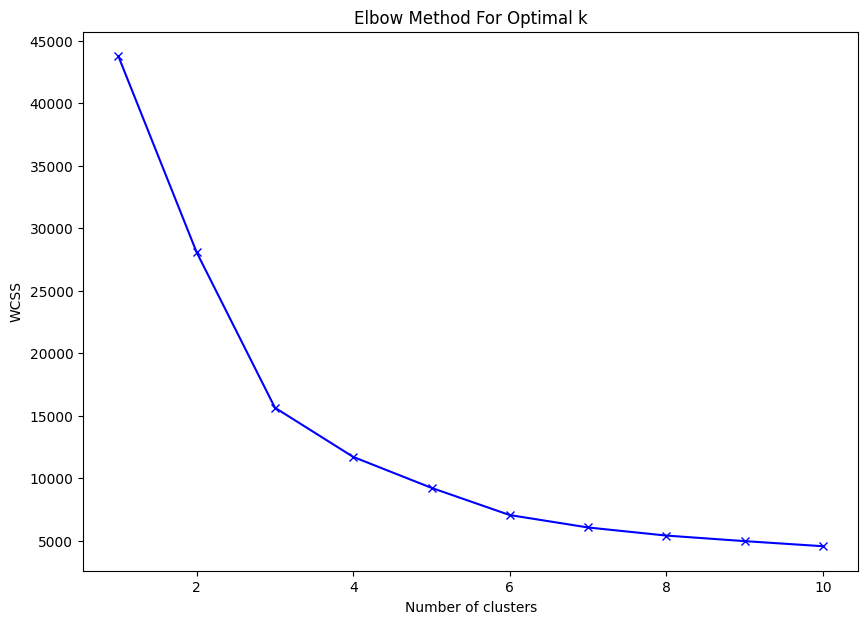
## در این قسمت تمامی داده ها با یک مقیاسگر (scaler ) ، به یک مقیاس (scale) نرملایز شده اند.

## پیدا کردن k بهینه

یکی از راه های پیدا کردن تعداد خوشه ها که ورودی الگوریتم k-means است ، روش elbow است.

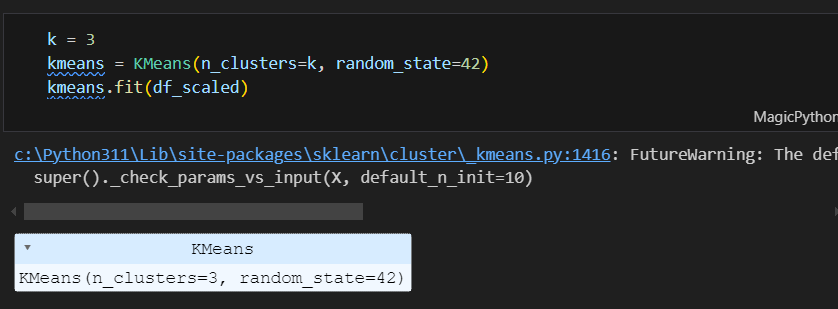


در این قسمت متود elbow پیاده سازی شده ، چندین بار خوشه بندی انجام شده و مجموع مربعات فاصله داده های خوشه ها برای هر تعداد از خوشه های خوشه بندی در wcss ذخیره شده و نمودار elbow رسم شده .

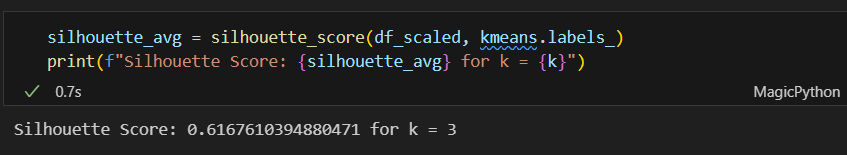


مشخص است که 3 الی 5 گزینه ها خوبی برای تعداد خوشه بندی چون که از انجا به بعد هر چه تعداد خوشه های خوشه بندی بیشتر میشود، مجموع مربعات فاصله داده های داخل خوشه های تغییر چندانی نمی کند که حاکی از تغییر نچندان مرکز خوشه هاست.

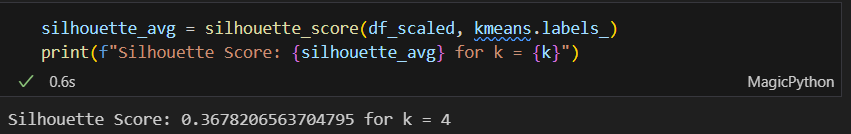
## الگوریتم k-means

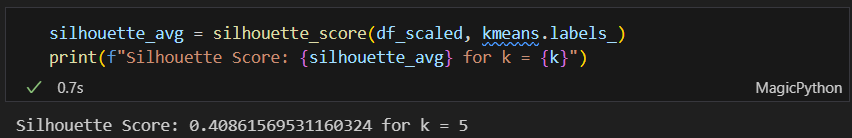
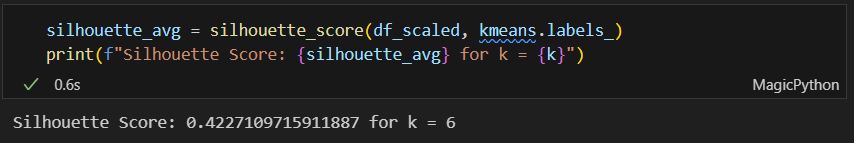
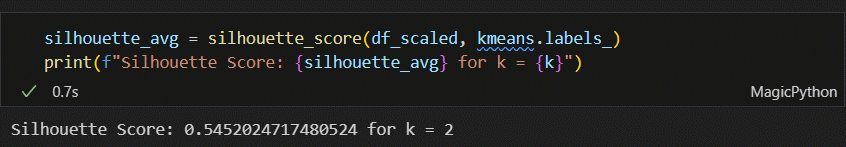
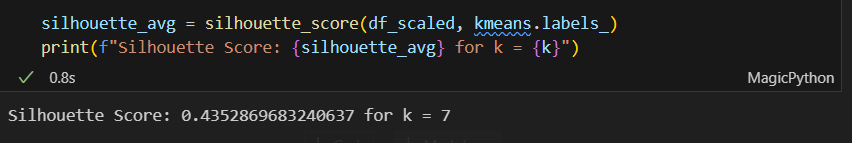


در این قسمت الگوریتمk-means رو داده ها با ورودی k=3 (۳ عدد خوشه مورد نظر) اجرا شده است.



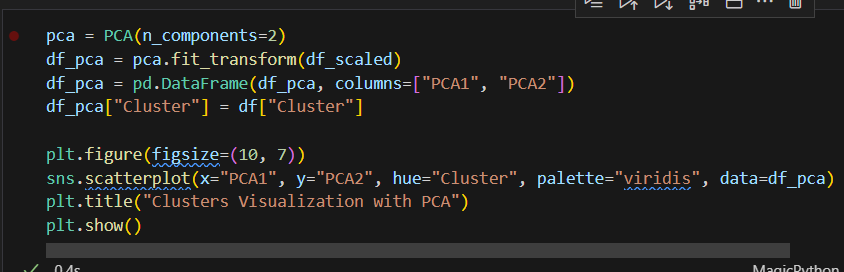
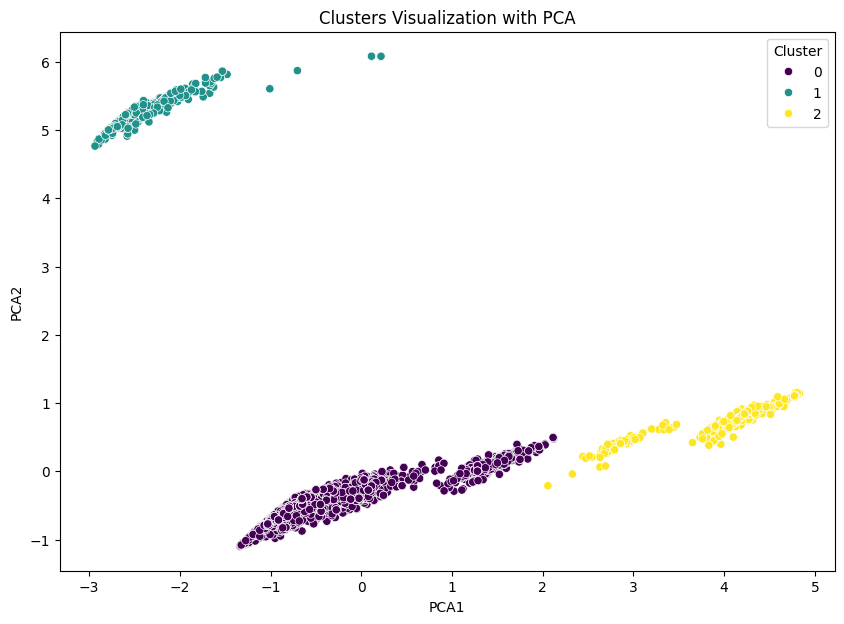
در این قسمت برای silhouette\_score میانگین خوشه ها به دست امده که این عدد نشان دهنده کیفیت خوشه بندی است . این فرایند برای تعداد خوشه های متفاوت انجام شده است .



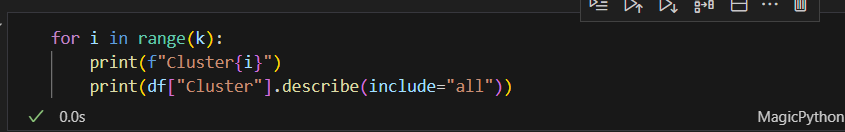
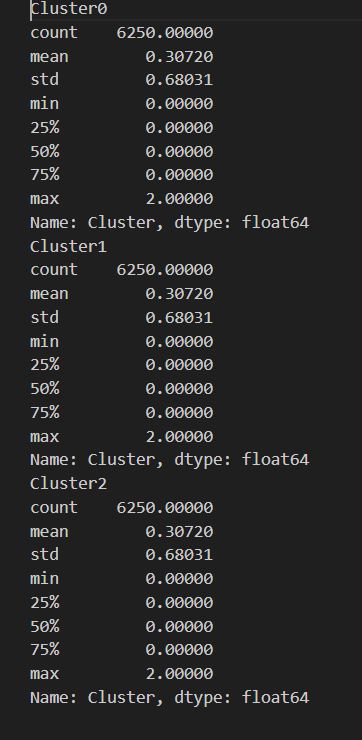
   

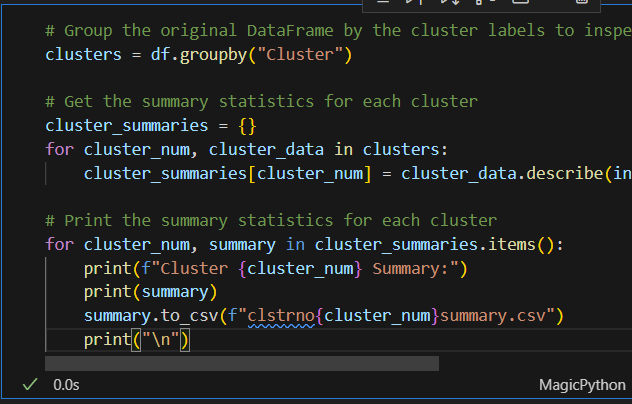
## نمایش داده ها

برای اینکه ابعاد و صفات داده های بیشتر از ۳ است از PCA استفاده شده تا نمایشی ۲بعدی در یک scatter-plot بتوان نمایش داد و بررسی کرد.

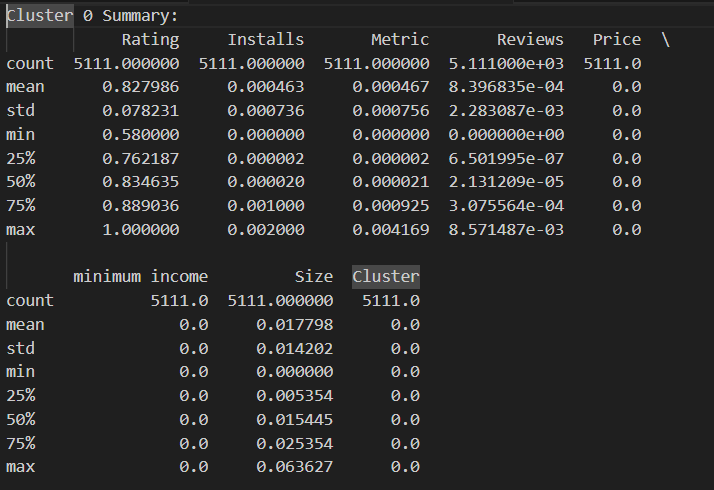
 

# تحلیل داده های نهایی

s 



### مشخصات خوشه اول

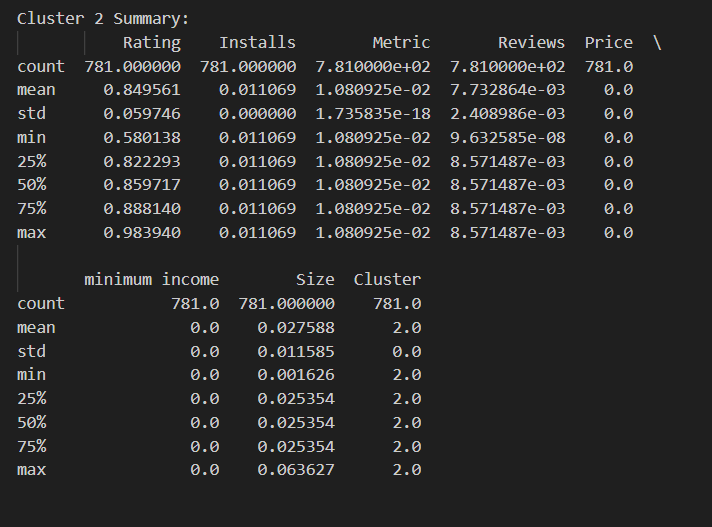


### مشخصات خوشه دوم

### 

در نگاه اول نکته متمایز کننده این خوشه ، درآمد زایی اپ های داخل ان است.

### مشخصات خوشه سوم



این خوشه شباهت به خوشه اول دارد ولی نکته متمایز کننده آن بهتر بودن metric و reviews آن ها میباشد که به طور میانگین این خوشه از لحاظ reviews 10 برابر بهتر و لحاظ metric کیفیت اپ که بر اساس داده ها ساهته شده است حدود 25 برابر بهتر است.

### بررسی نهایی

به نظر میرسد که خوشی بندی ها اپ ها را به دسته درامد زا ، بدون درامد-بدون کیفیت و بدون درامد-باکیفیت دسته بندی میکنند.