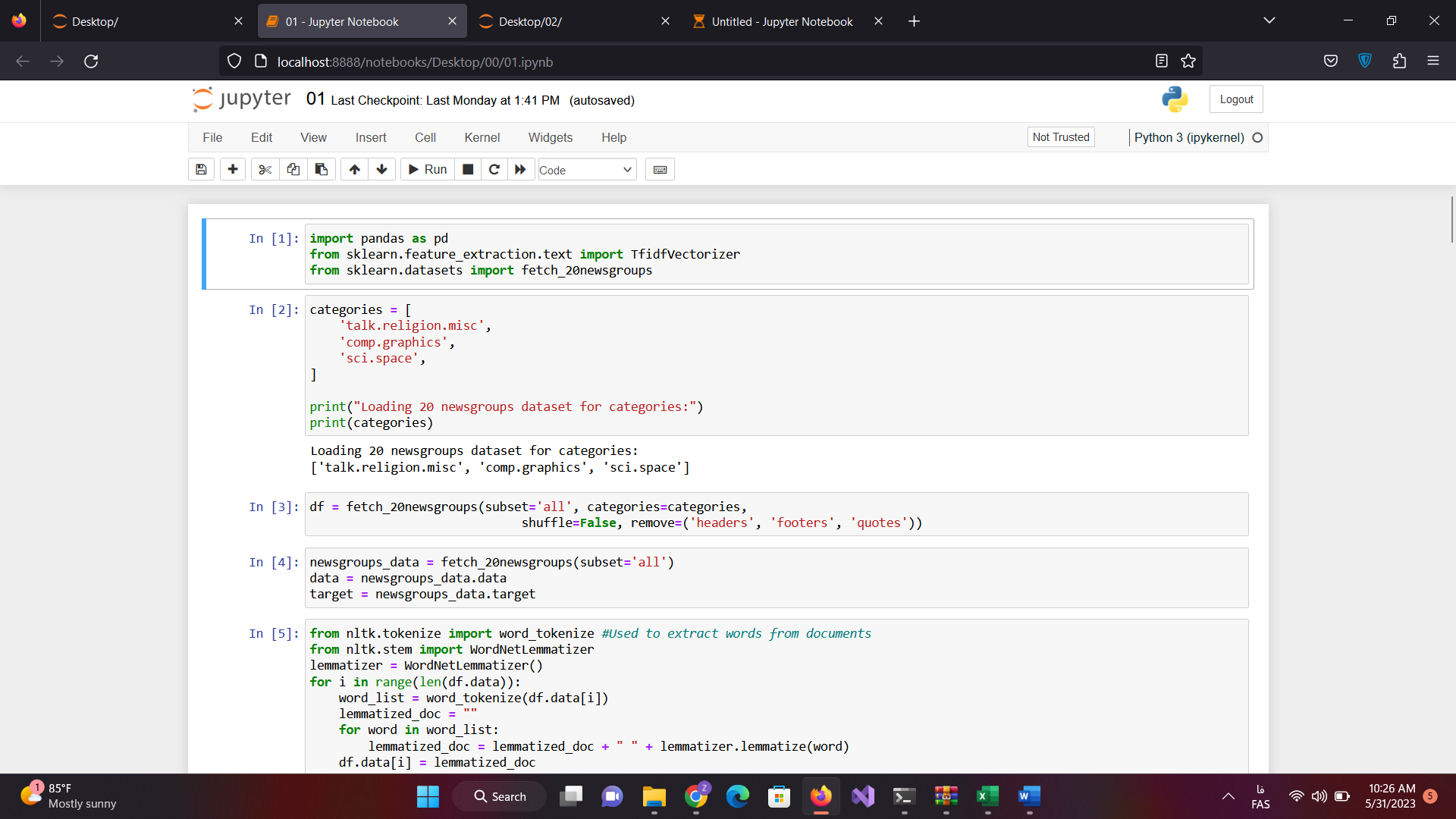
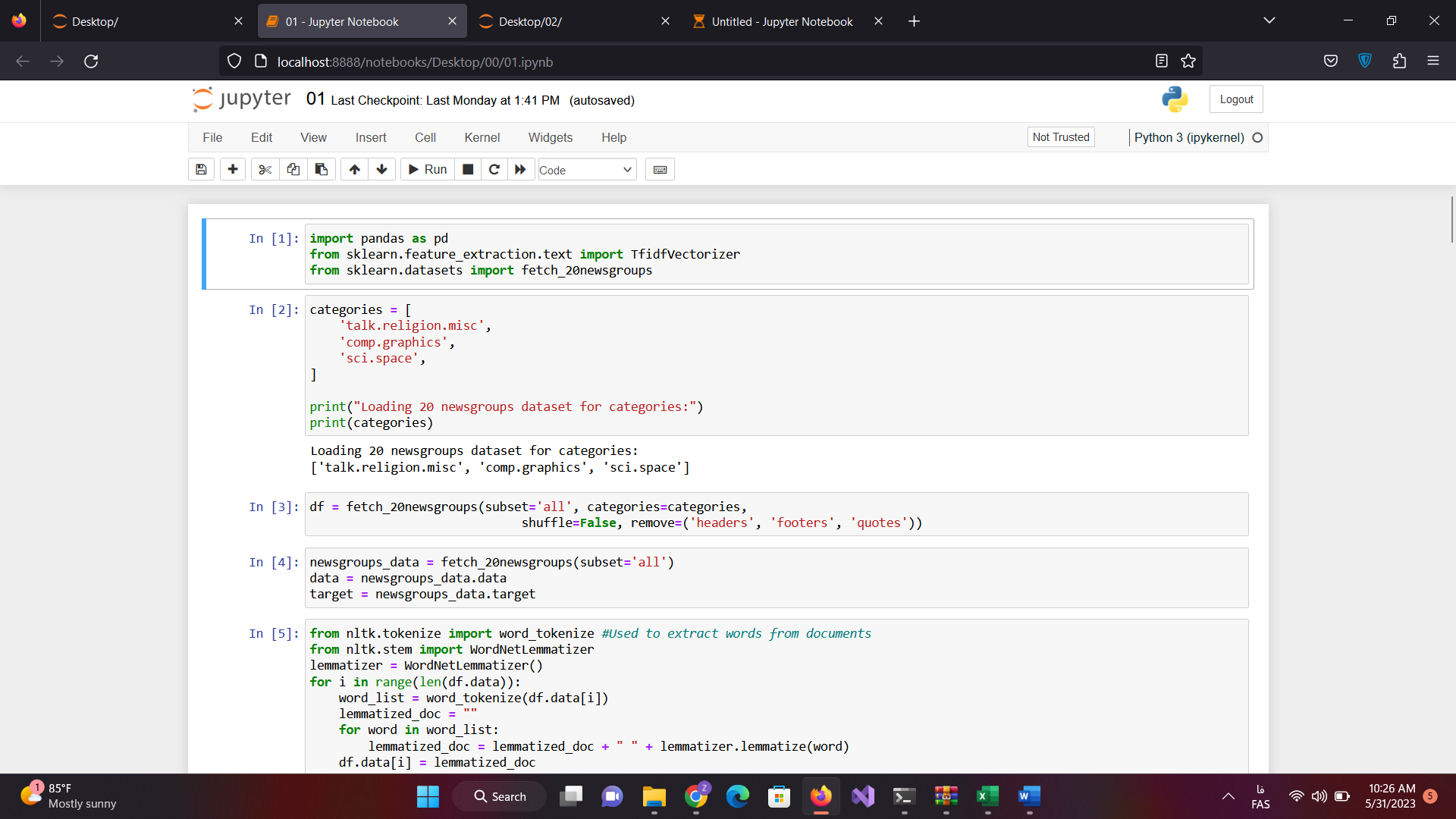
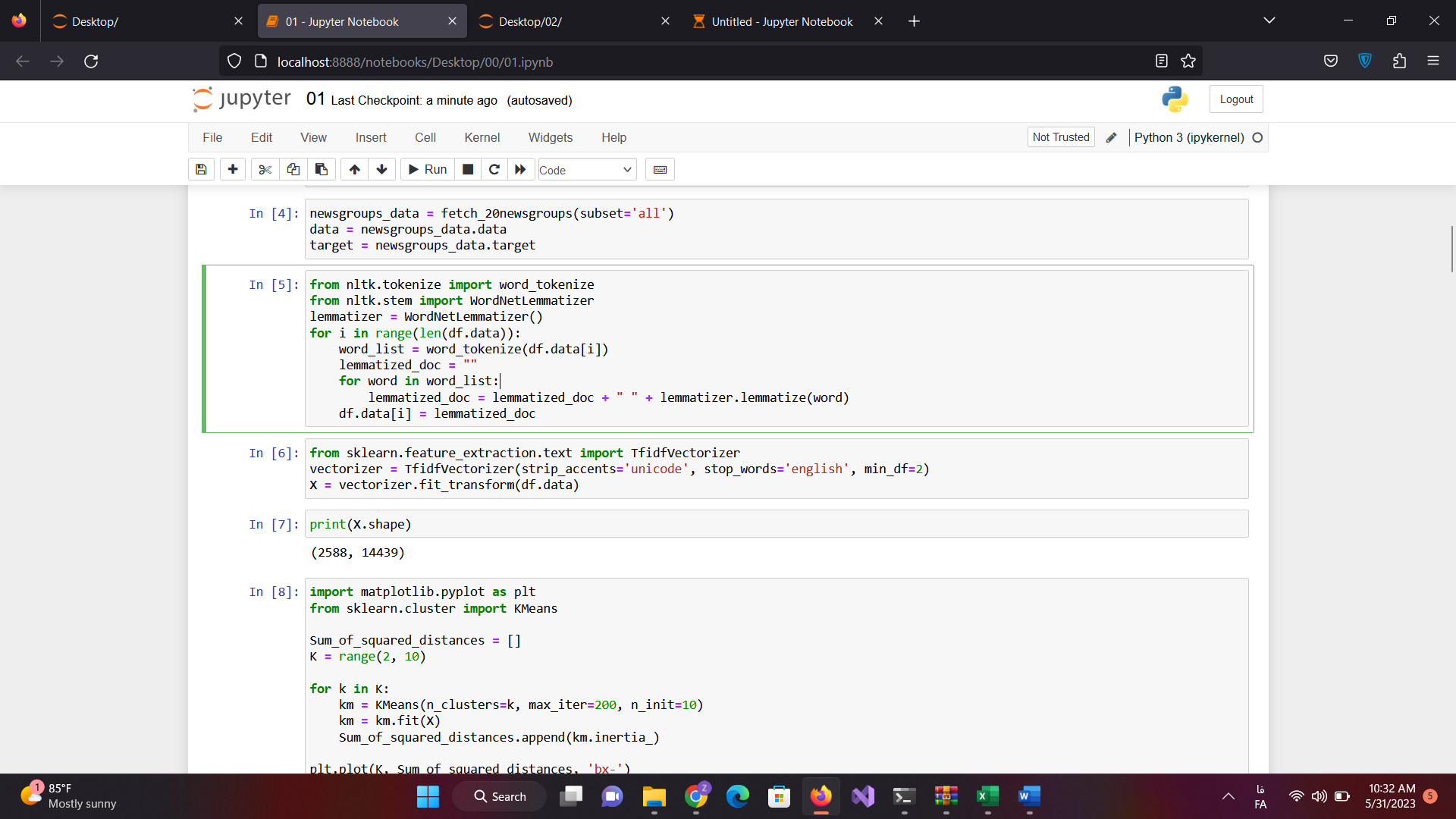
در قسمت کلاستر ما آمده ایم یک کتابخونه به اسم fetch\_20newsgroups که 20 تا از خبرجدید که با استفاده از categories که ما به ان میدهیم میگیرد و این داده کلاستر ما میشود.



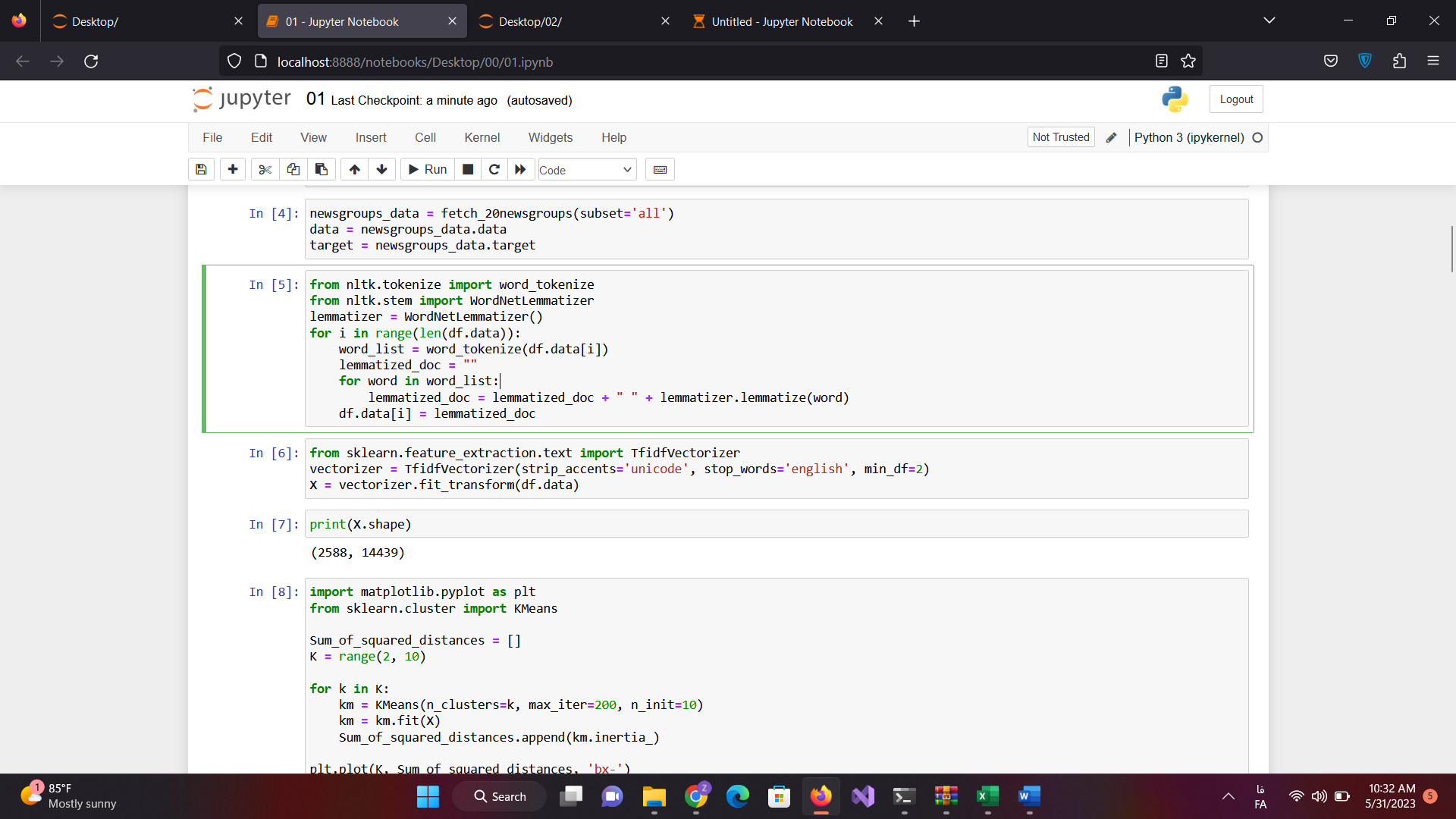
این یک داده با سرصفحه و قالب سایت میگیره که باید موارد اضافی را حذف کنیم تا داده اصلی را به ما بدهد. و داده اصلی و تارگت آن را جدا میکنیم که البته توی خوشه بندی زیاد موضوع کاربردی ندارد.



از تابع زیر استفاده میکنیم تا پیش پردازش روی متن را انجام بدهیم اول می آید کلمات را توکن توکن میکنه و بعد با WordNetLemmatizer برای واژه سازی استفاده می شود که فرآیند کاهش کلمات به شکل پایه یا ریشه آنها است. که در اینجا کلمات کمتر میشه و پردازش ما اسان تر میشود.

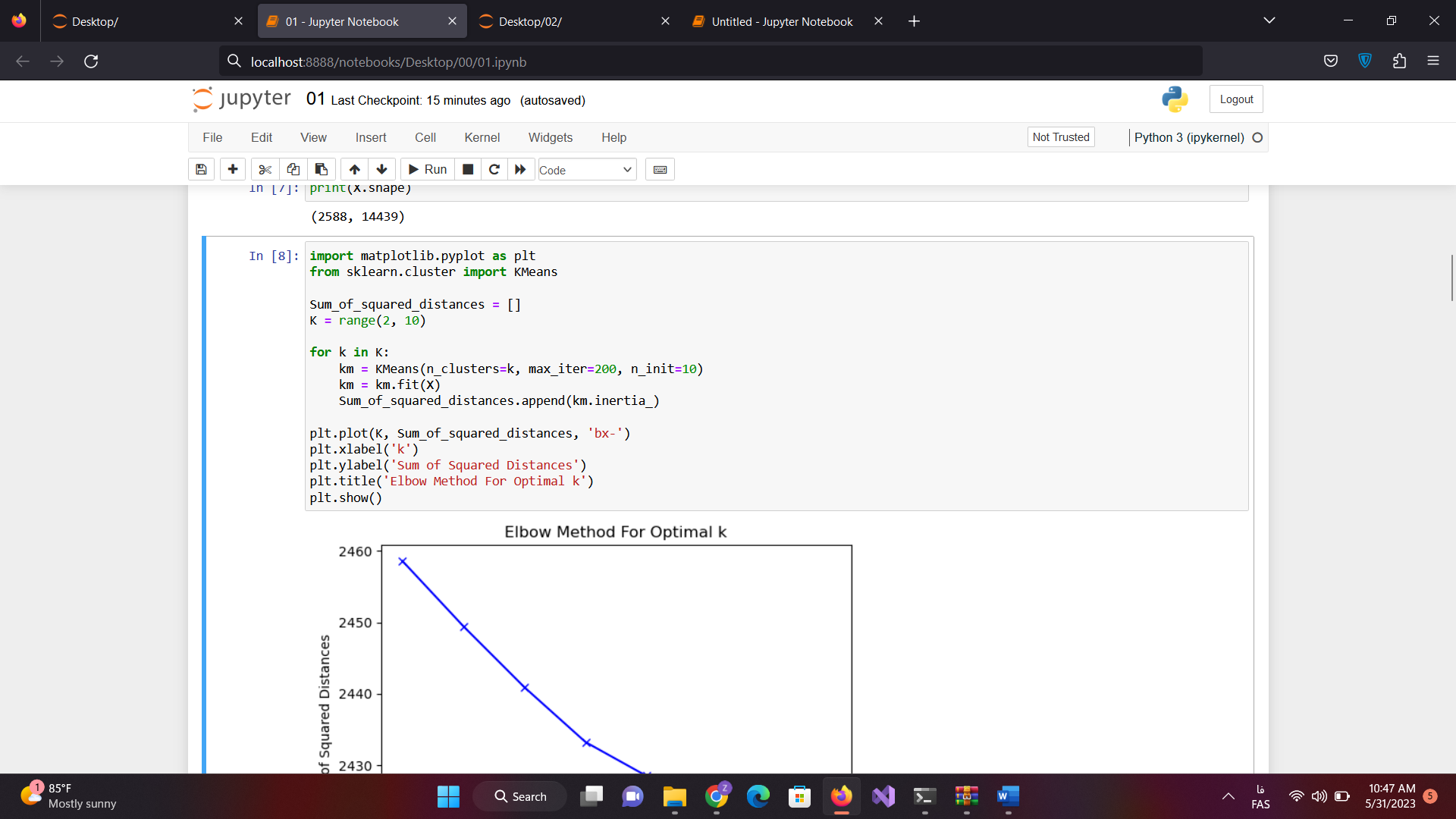


حال میایم بر استاپ ورد ها را حذف میکنیم و بر روی آن tf-idf میزنیم. میبینیم که shape ما چند در چند را به ما نشان میدهد.

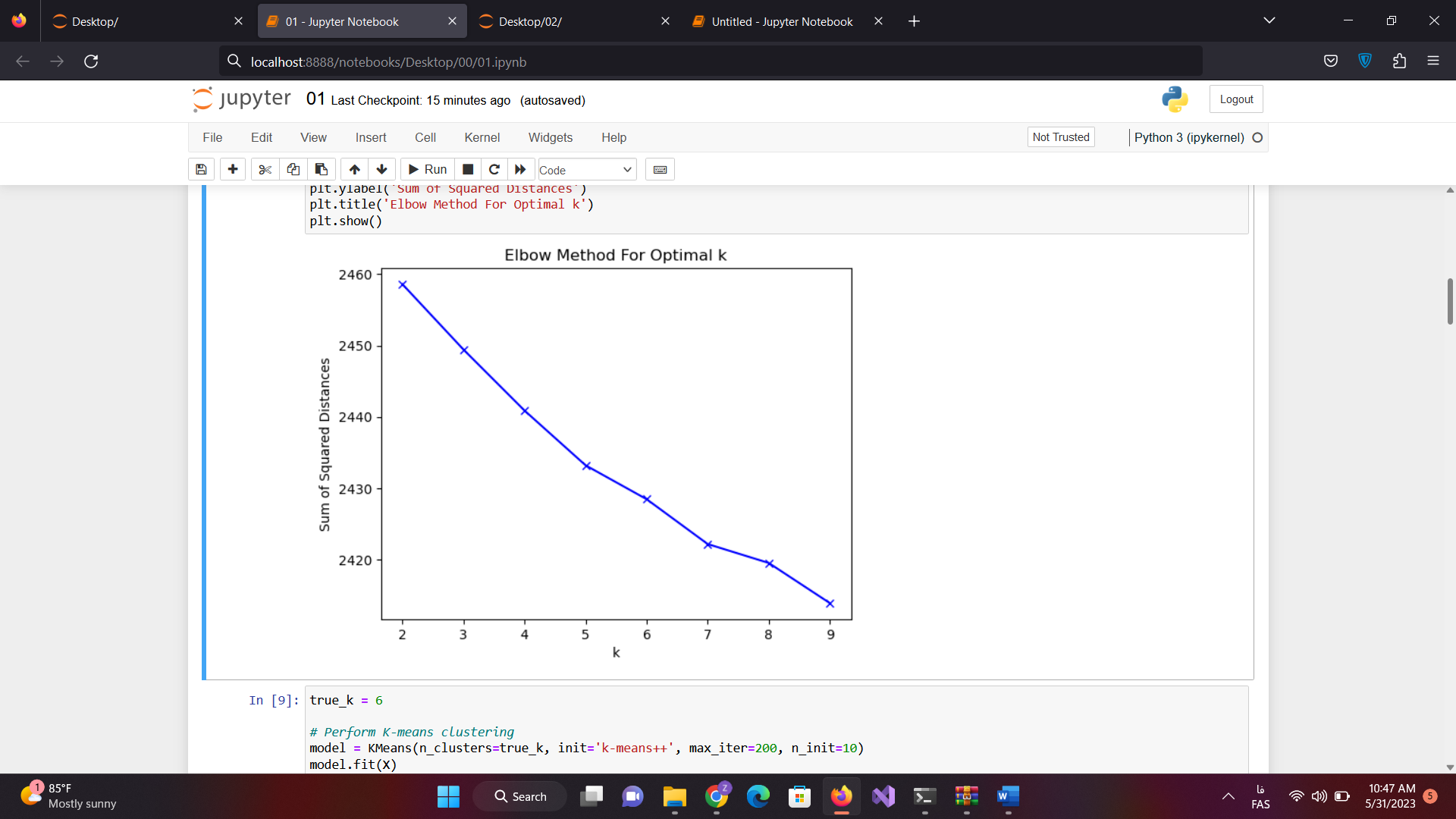


هر ردیف از متغیر X نمایش برداری از یک صفحه خبر است. از این رو، می توانیم از X به عنوان ورودی برای الگوریتم k-means استفاده کنیم.

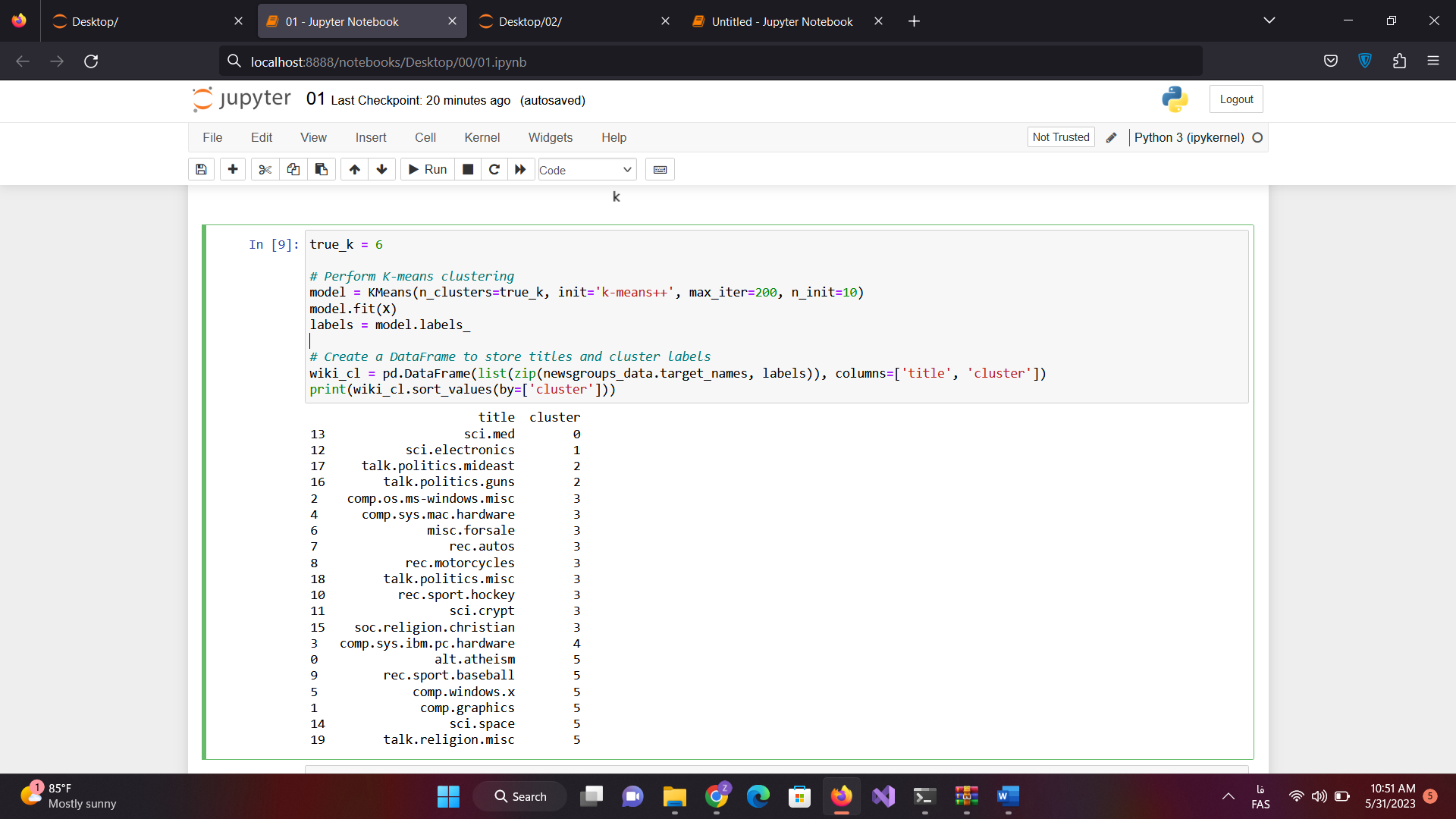
ابتدا باید در مورد تعداد خوشه ها تصمیم گیری کنیم. در اینجا از روش  [elbow method](https://en.wikipedia.org/wiki/Elbow_method_(clustering)) استفاد خواهیم کرد.



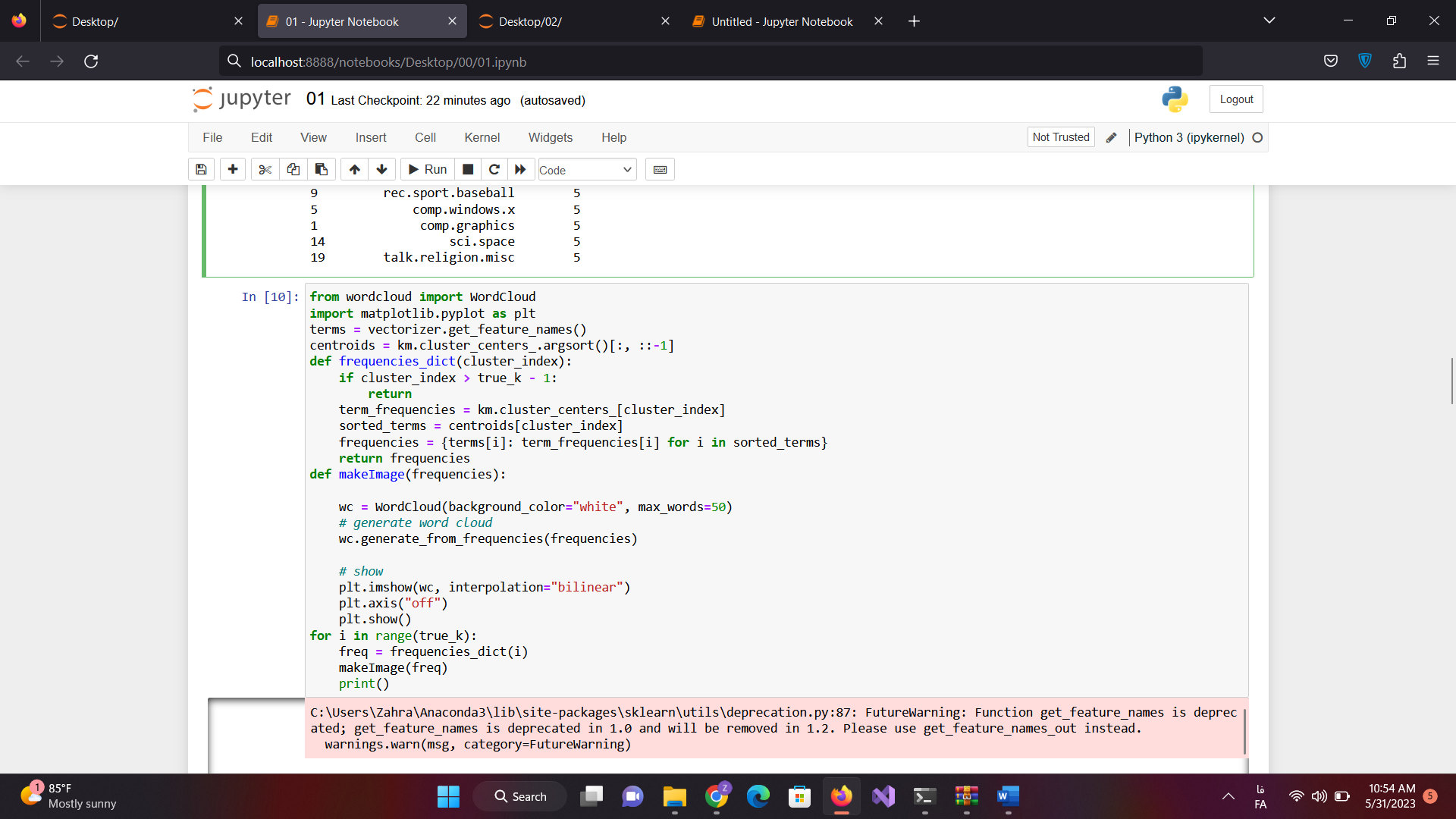
خروجی آن به صورت زیر است.

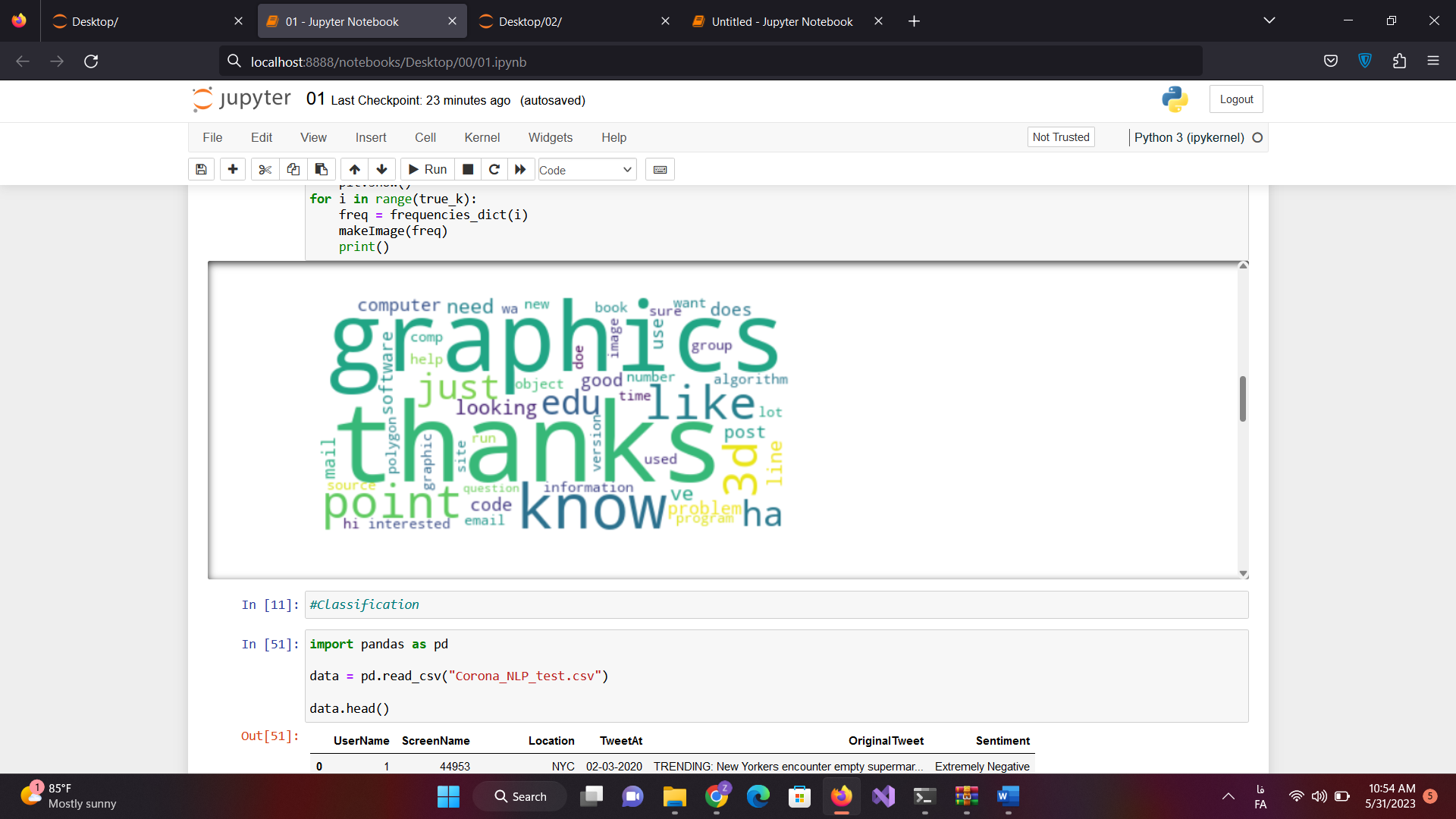
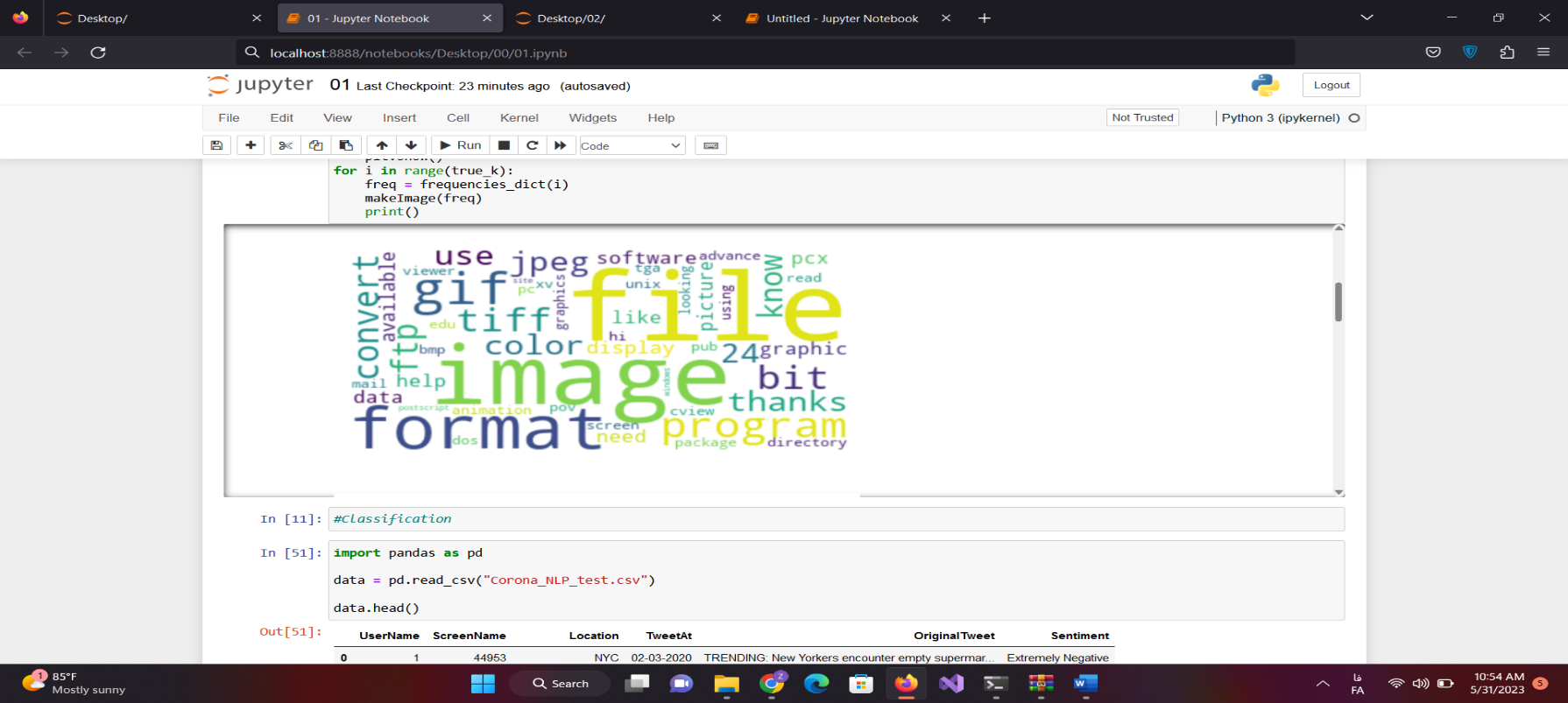


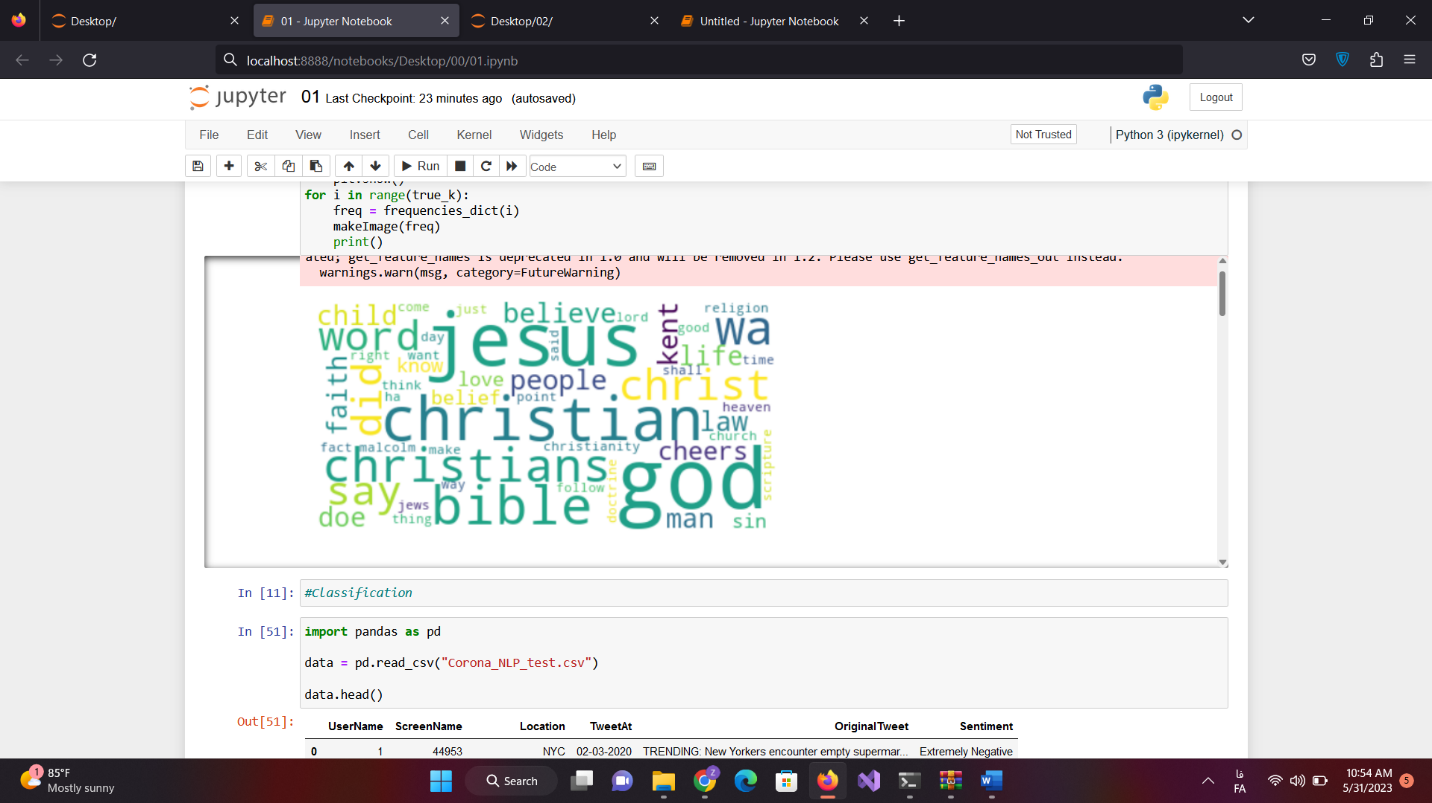
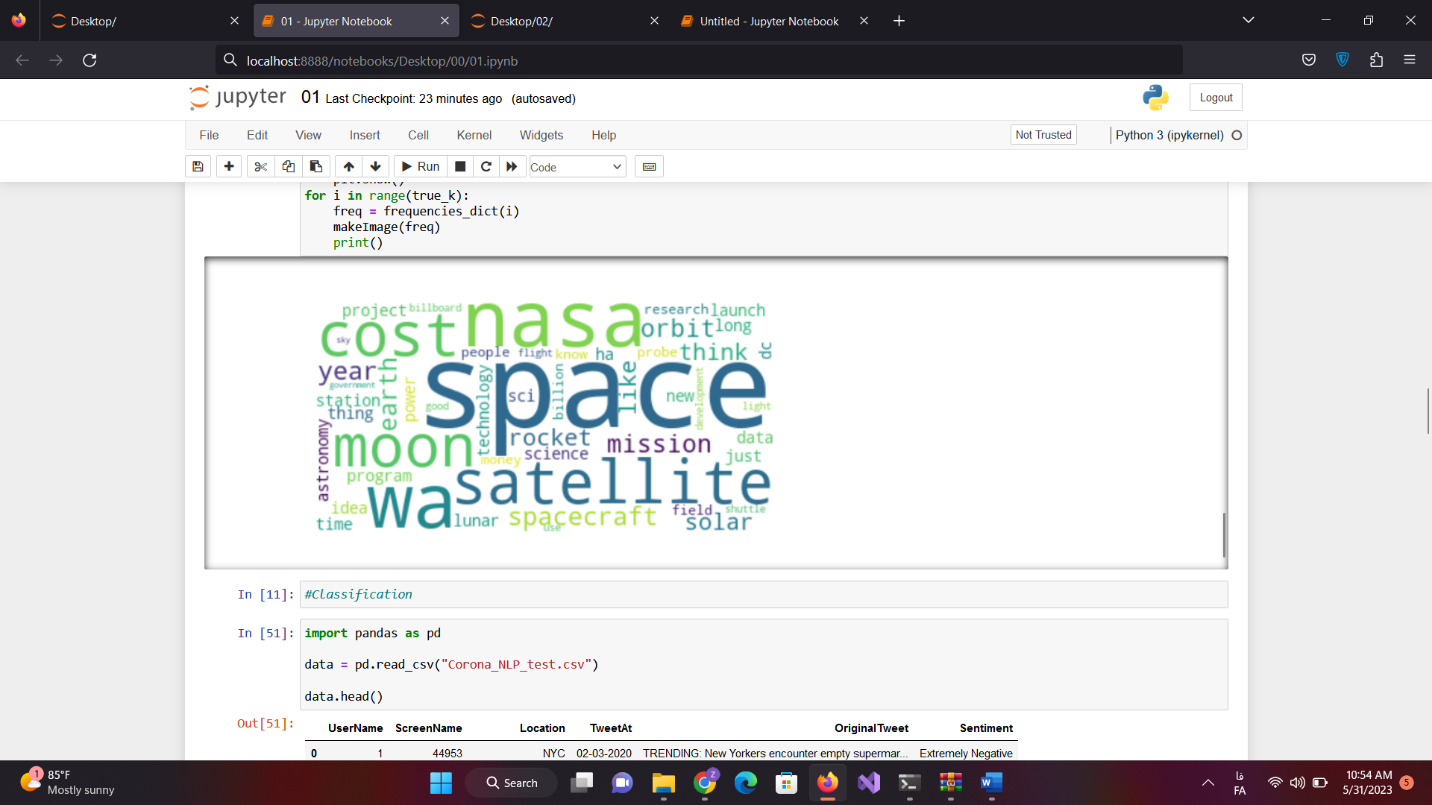
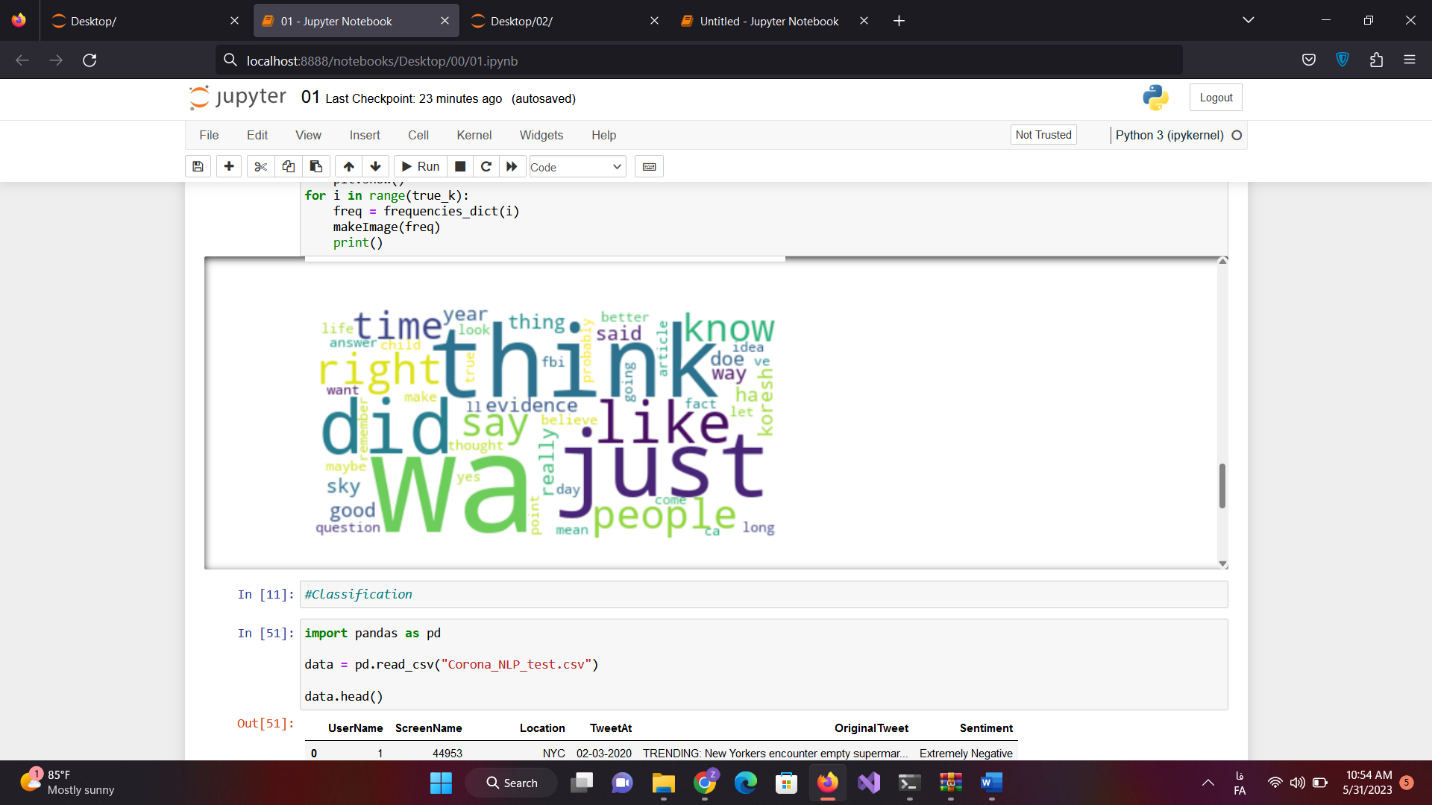
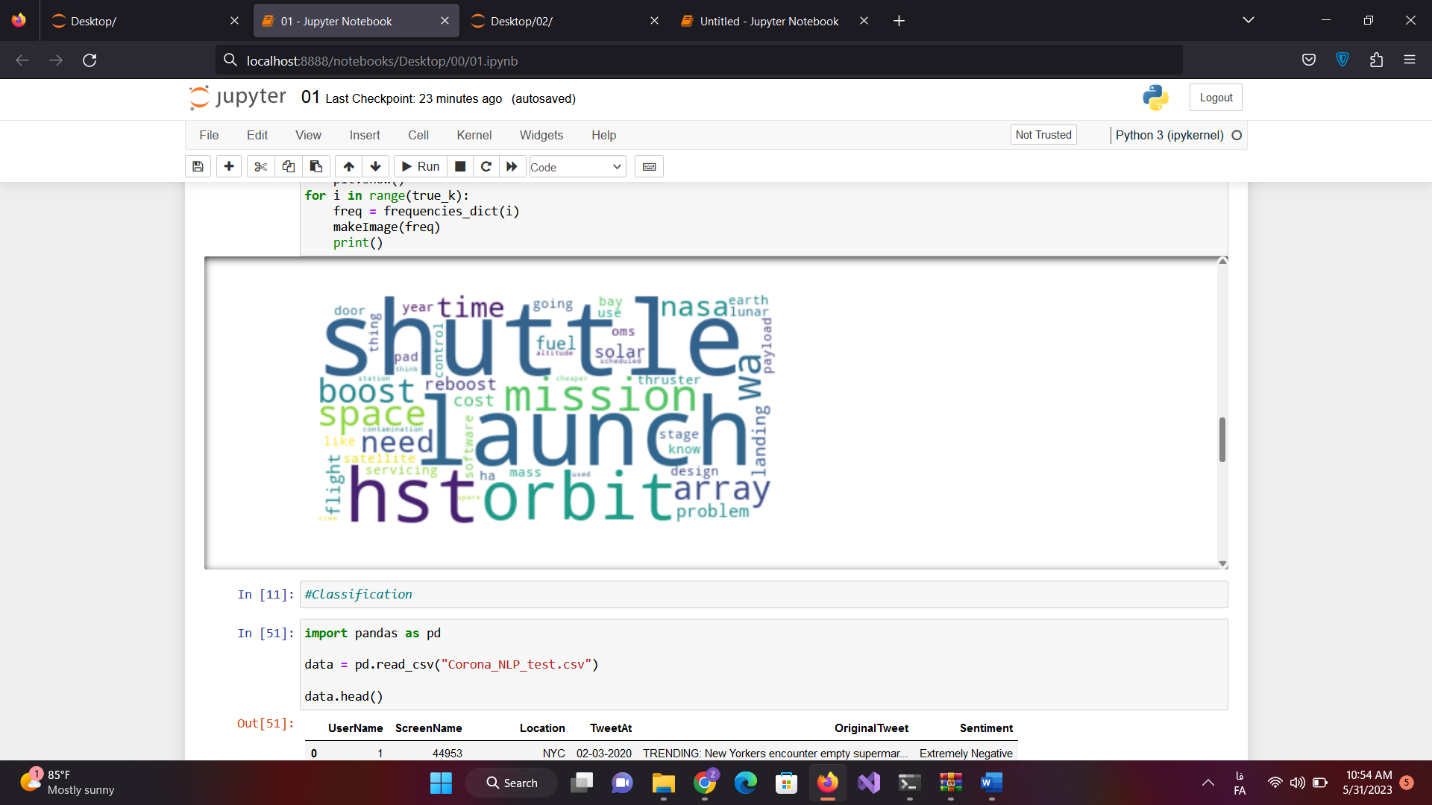
و سپس خوشه بندی 6 در نظر میگیریم. و خوشه بندی به روش k-means را انجام میدهیم و میبینیم که این مدل خوشه بندی را به چ صورتی برای ما انجام میدهد.



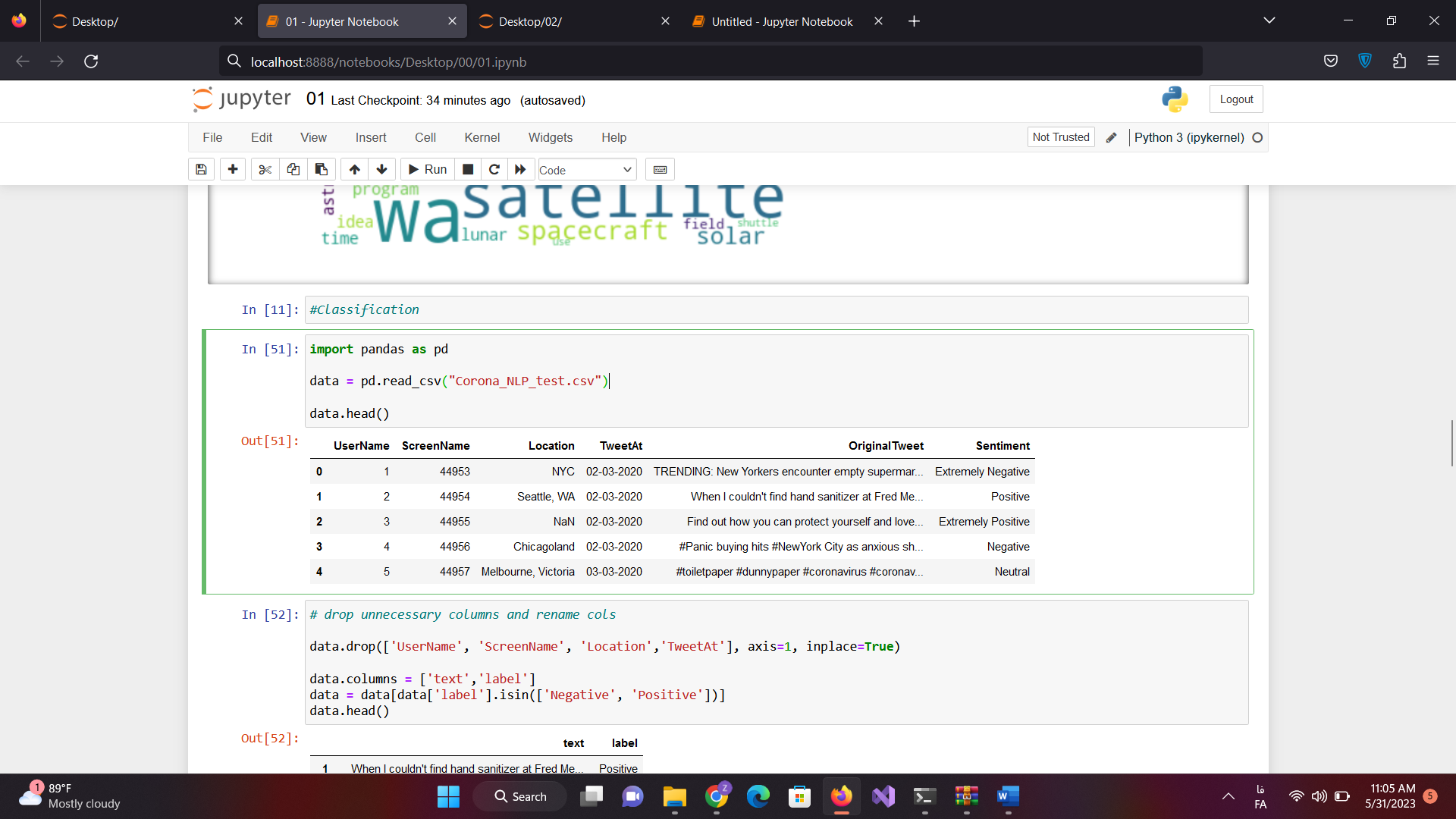
از روش ابرکلمه استفاده میکنیم تا به ما خروجی را به صورت شکل نشان دهد تا جزئیات بیشتری بتوانیم بدست اوریم.



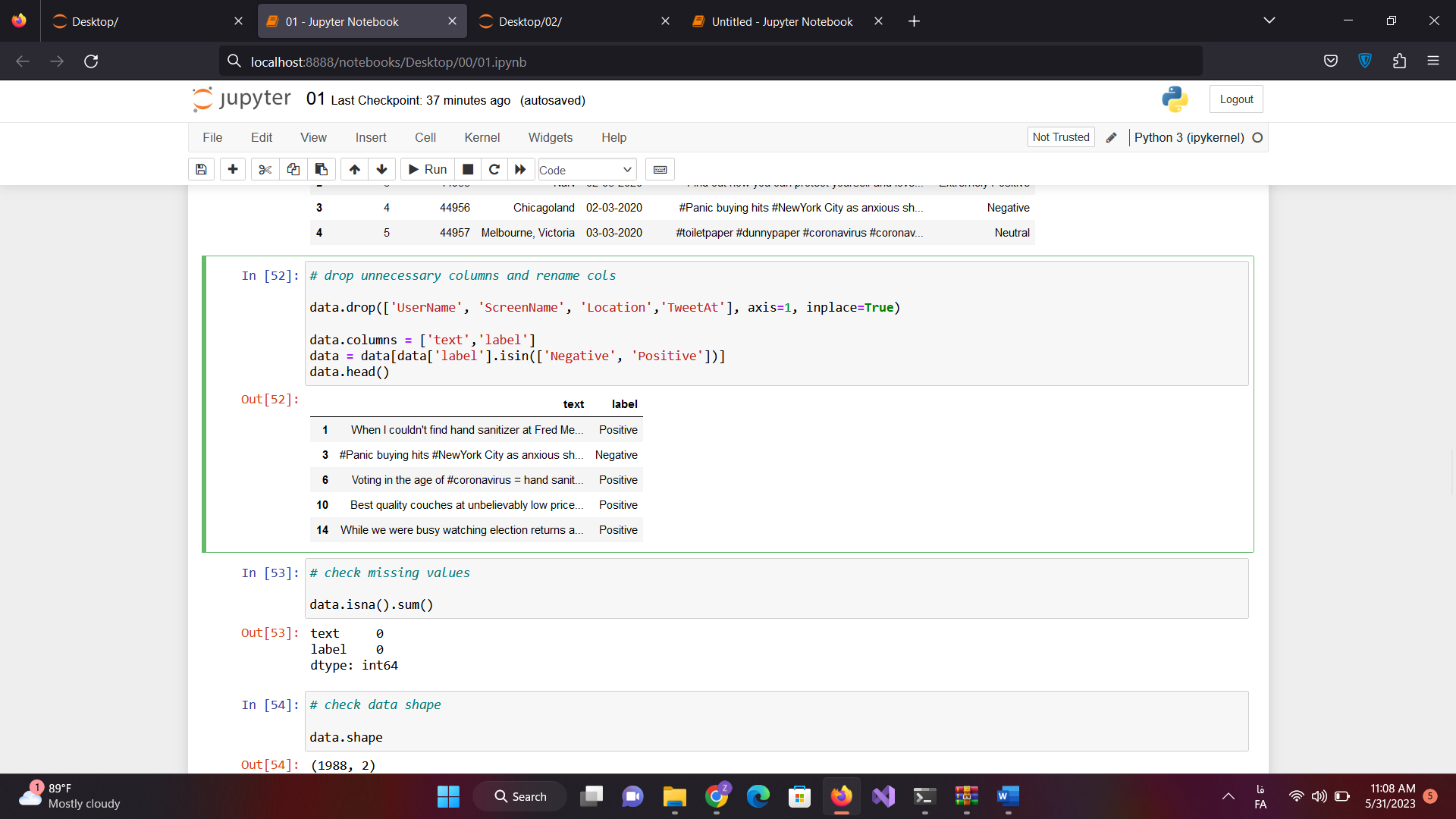




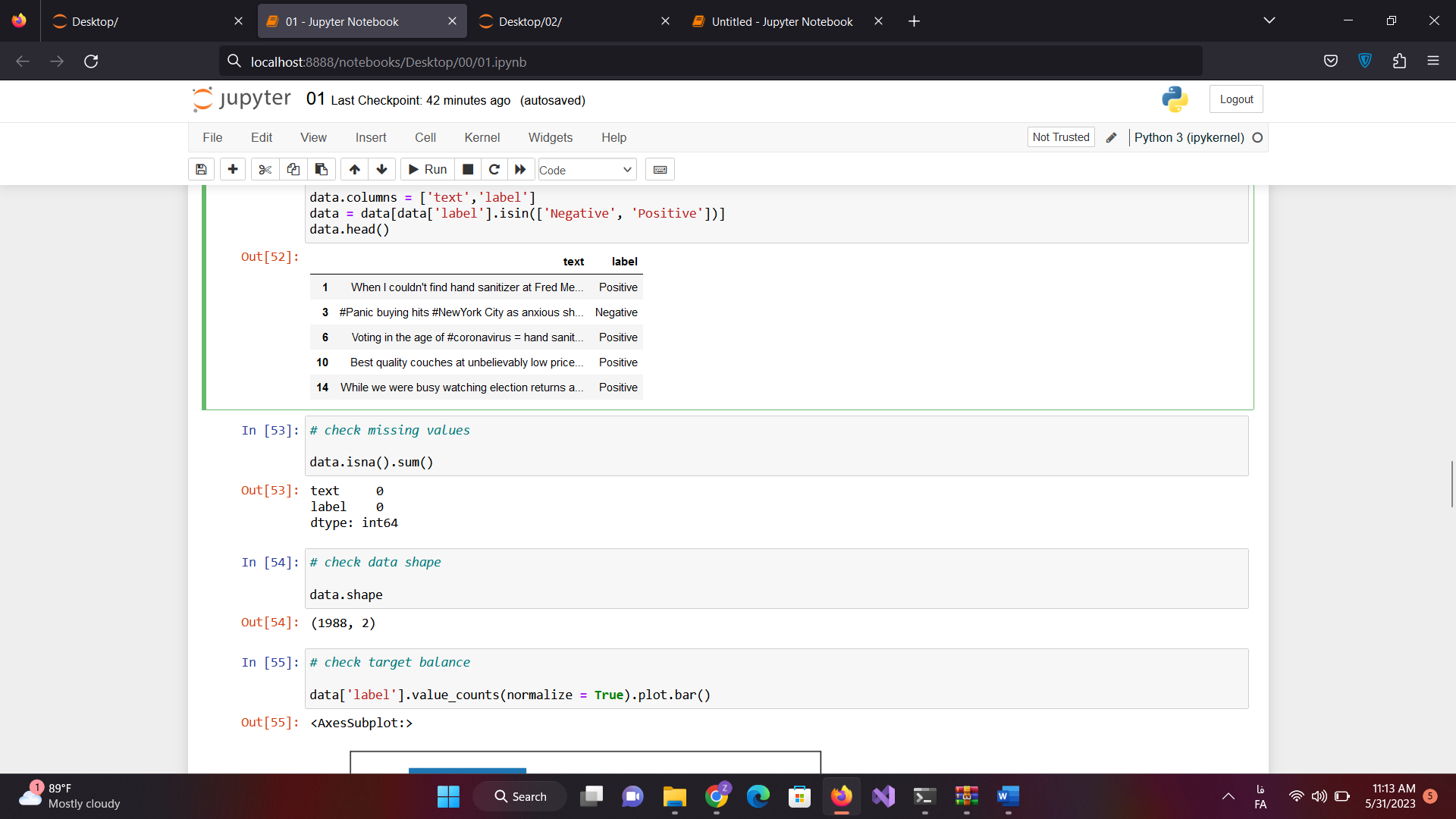
بعدی درباره کلاس بندی هست که ما یک دیتاست پیدا کردیم که درباره تویت هایی که یک جامعه زده و اینک جواب کروناشون مثبت شده یا نه و بر روی این داده کار میکنیم.



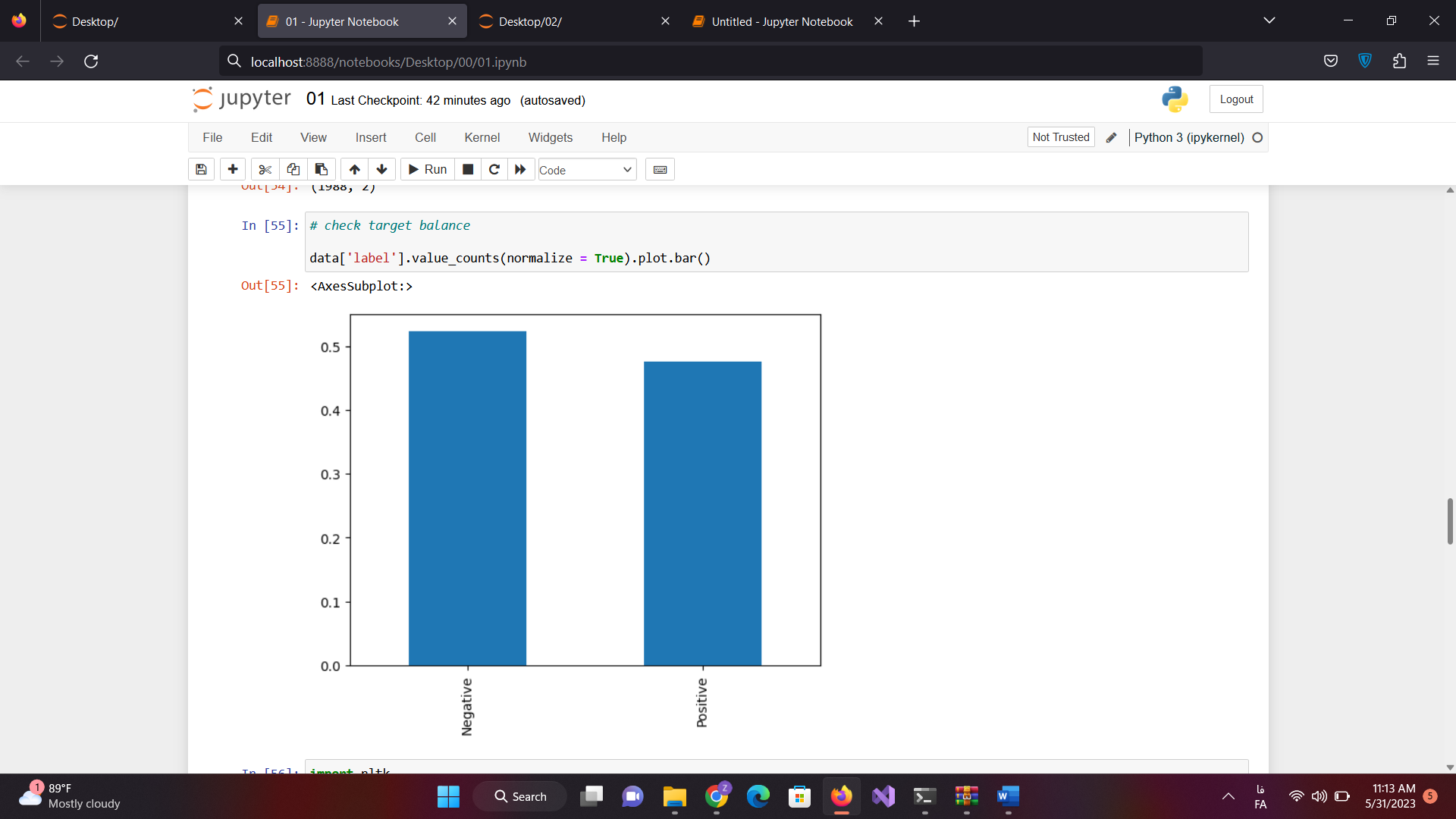
که در اینجا ستون هایی اضافی را حذف میکنیم و فقط اورجینال توییت رو نگه میداریم و فقط بیماری که مشخصه که کامل جواب تست کروناشون مثبت یا منفی است.



میاییم مقداری EDA اساسی انجام می دهیم و میبینیم مقادیر گمشده در مجموعه داده وجود ندارد و تعادل هدف را میبینیم.



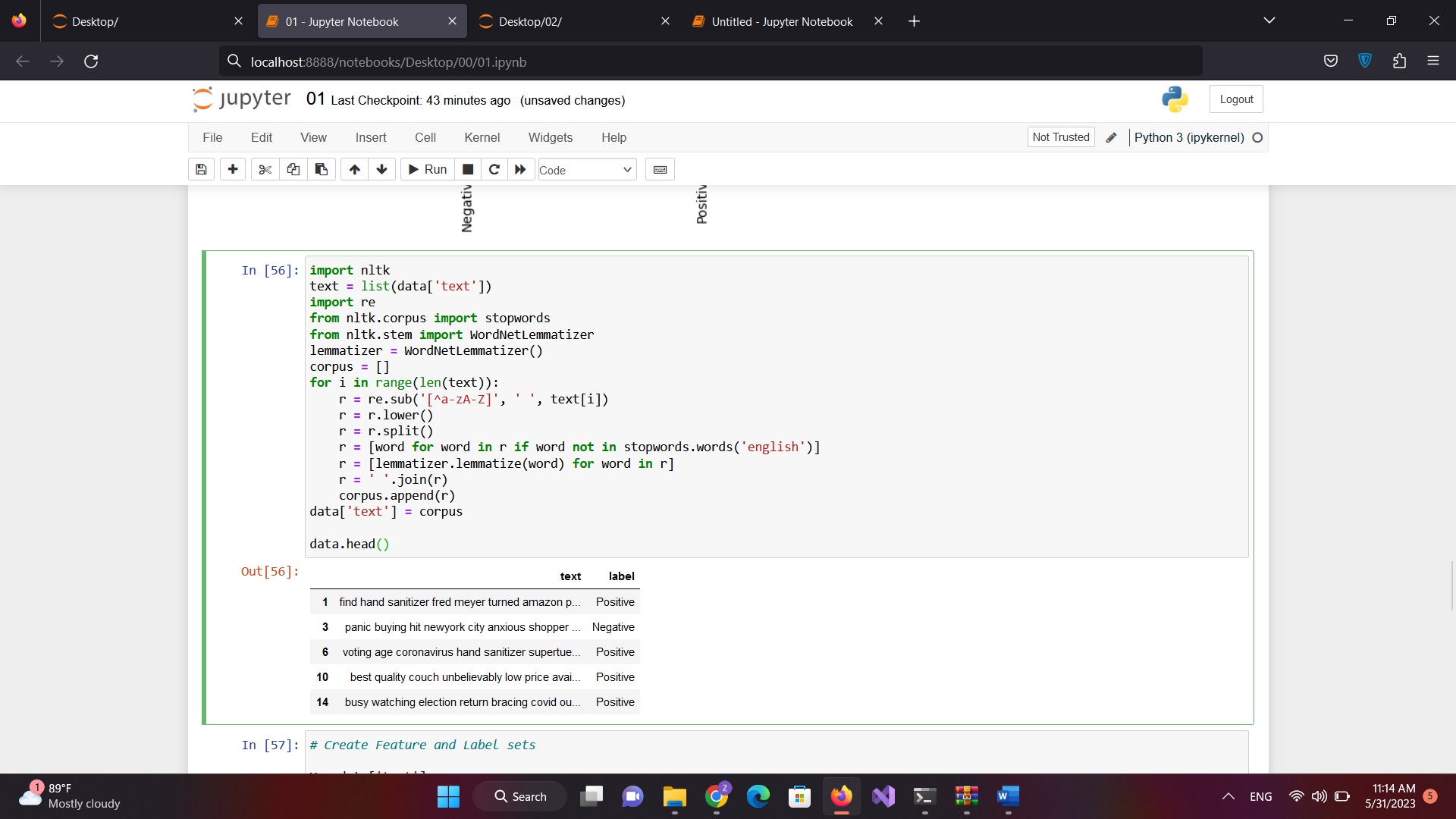
حال میایم میبنیم که چقدر از داده ها مثبت و منفی است plot.bar آن را رسم میکنیم که میبنیم تقریبا نرمال توزیع شده است.



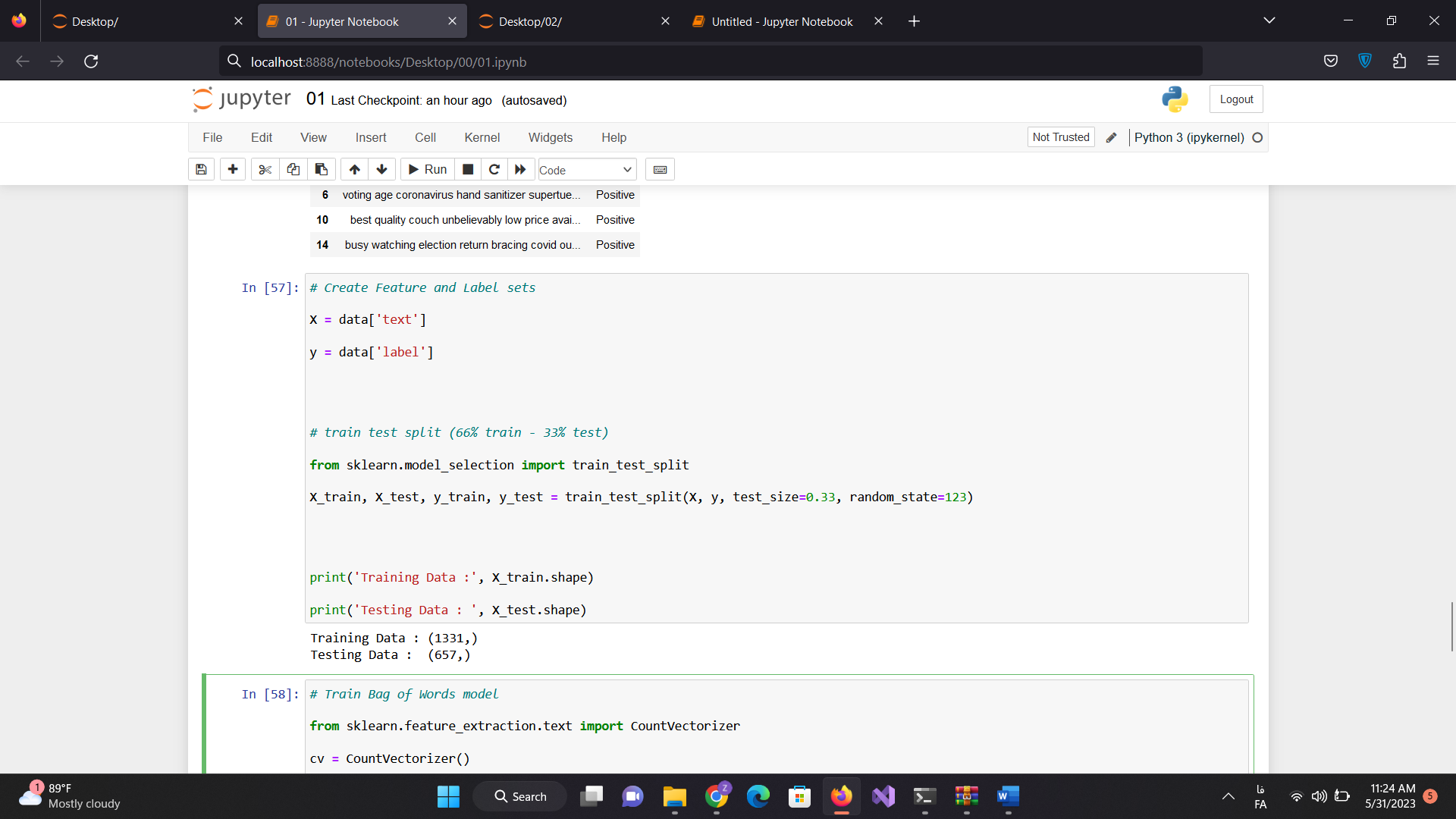
حال داده ها رو بر روی ان پیش پردازش انجام می دهیم.

* Remove all special characters
* Lowercase all the words
* Tokenize
* Remove stopwords
* Lemmatize

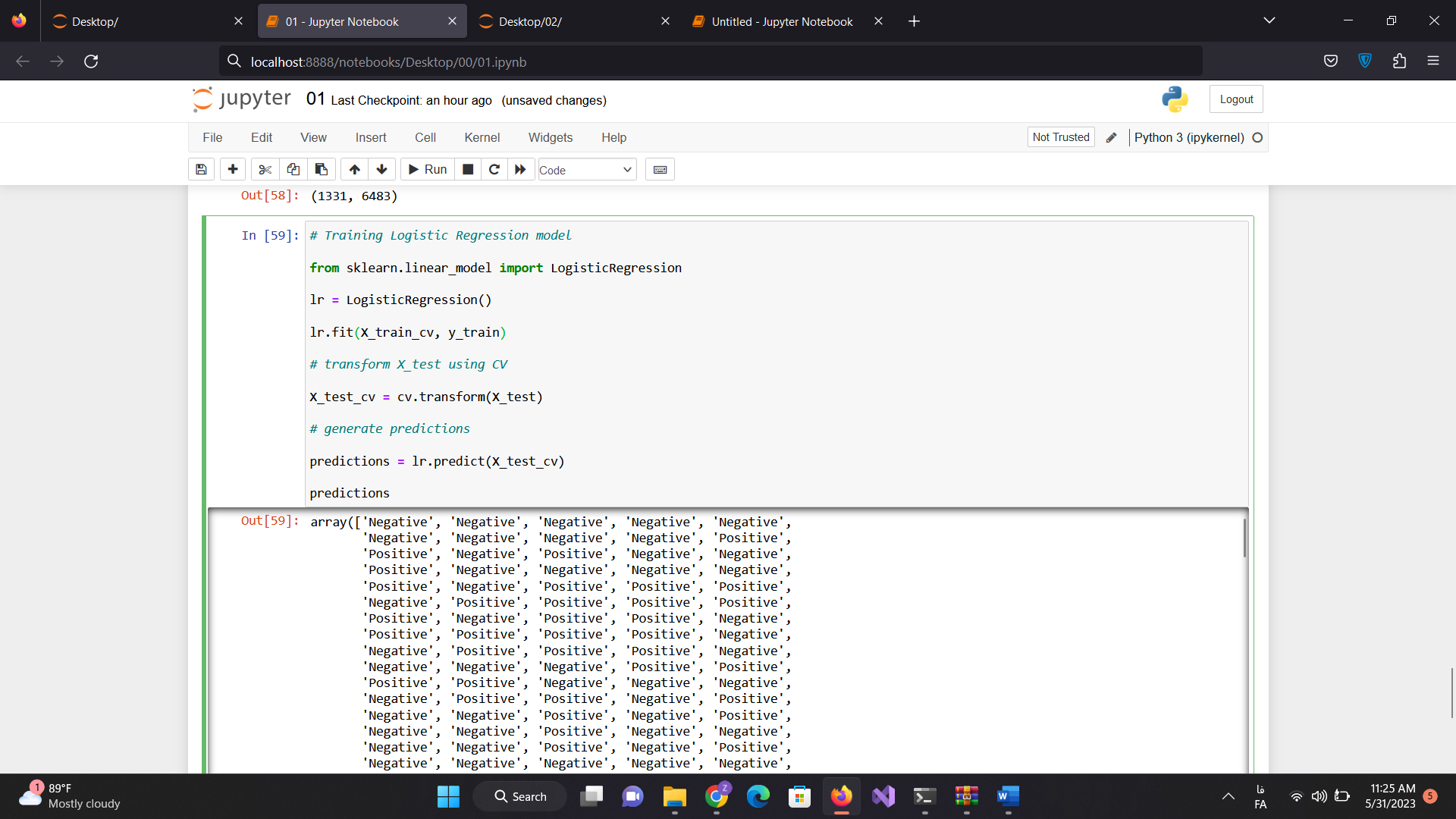
و این موارد را بر روی داده ها انجام میدهیم.



میاییم مجموعه داده را به آموزش تقسیم کنیم و قبل از استخراج ویژگی آزمایش کنیم.



در اینجا، از مدل Bag of Words (CountVetorizer) برای تبدیل متن پاک شده به ویژگی های عددی استفاده می کنیم. این برای آموزش مدل یادگیری ماشین مورد نیاز است.



در این قسمت به آموزش یک مدل رگرسیون لجستیک و ارزیابی ماتریس سردرگمی مدل آموزش‌دیده می‌پردازیم.

