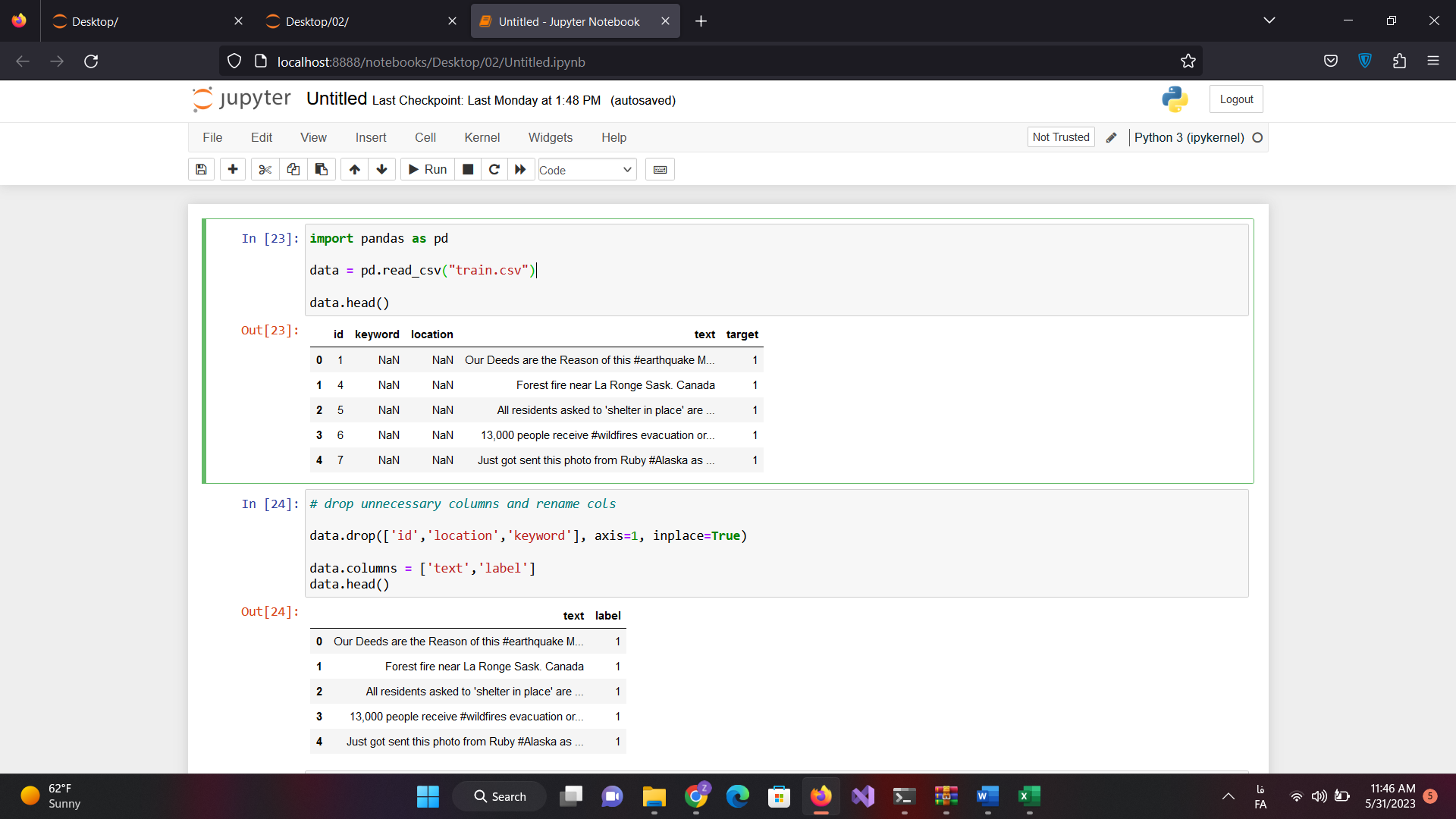
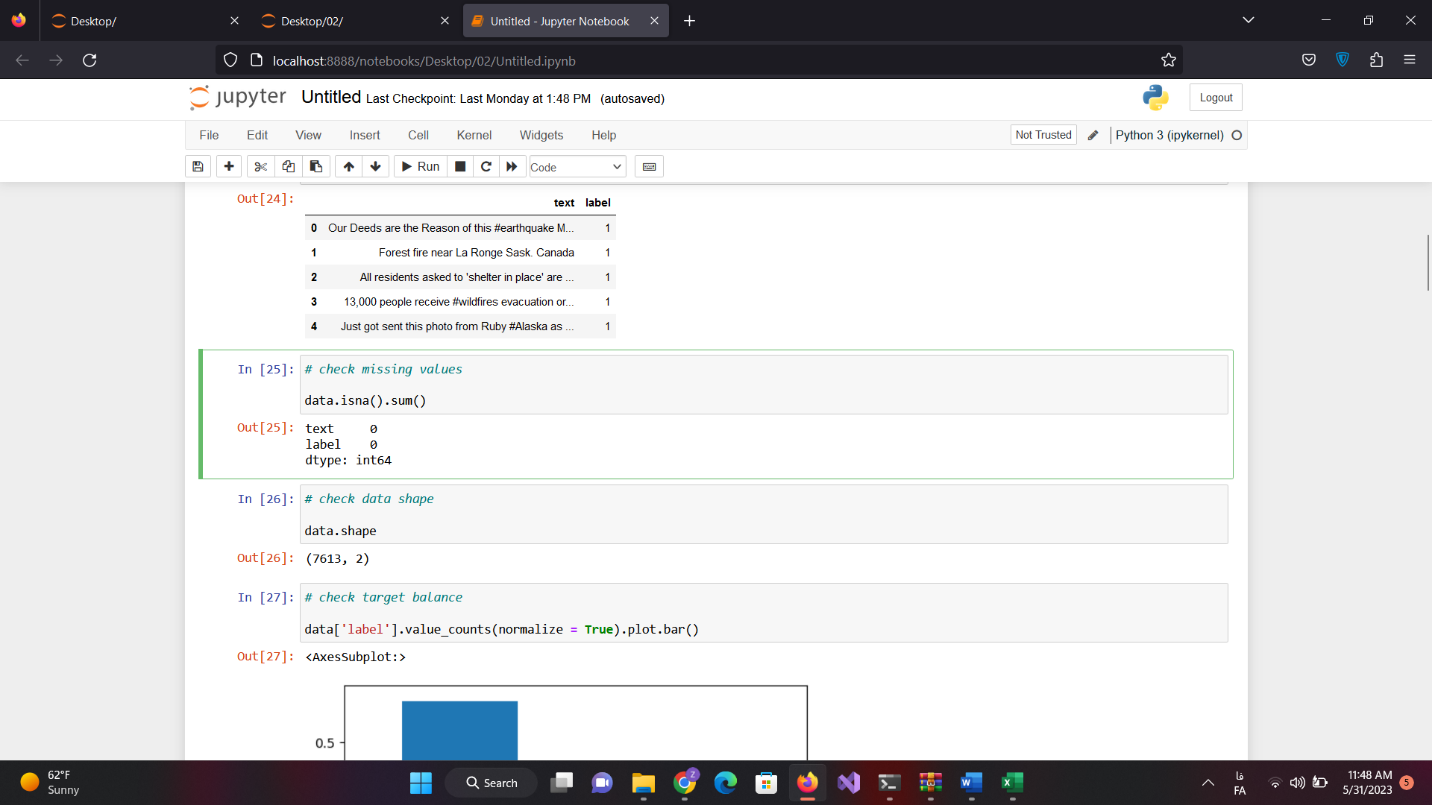
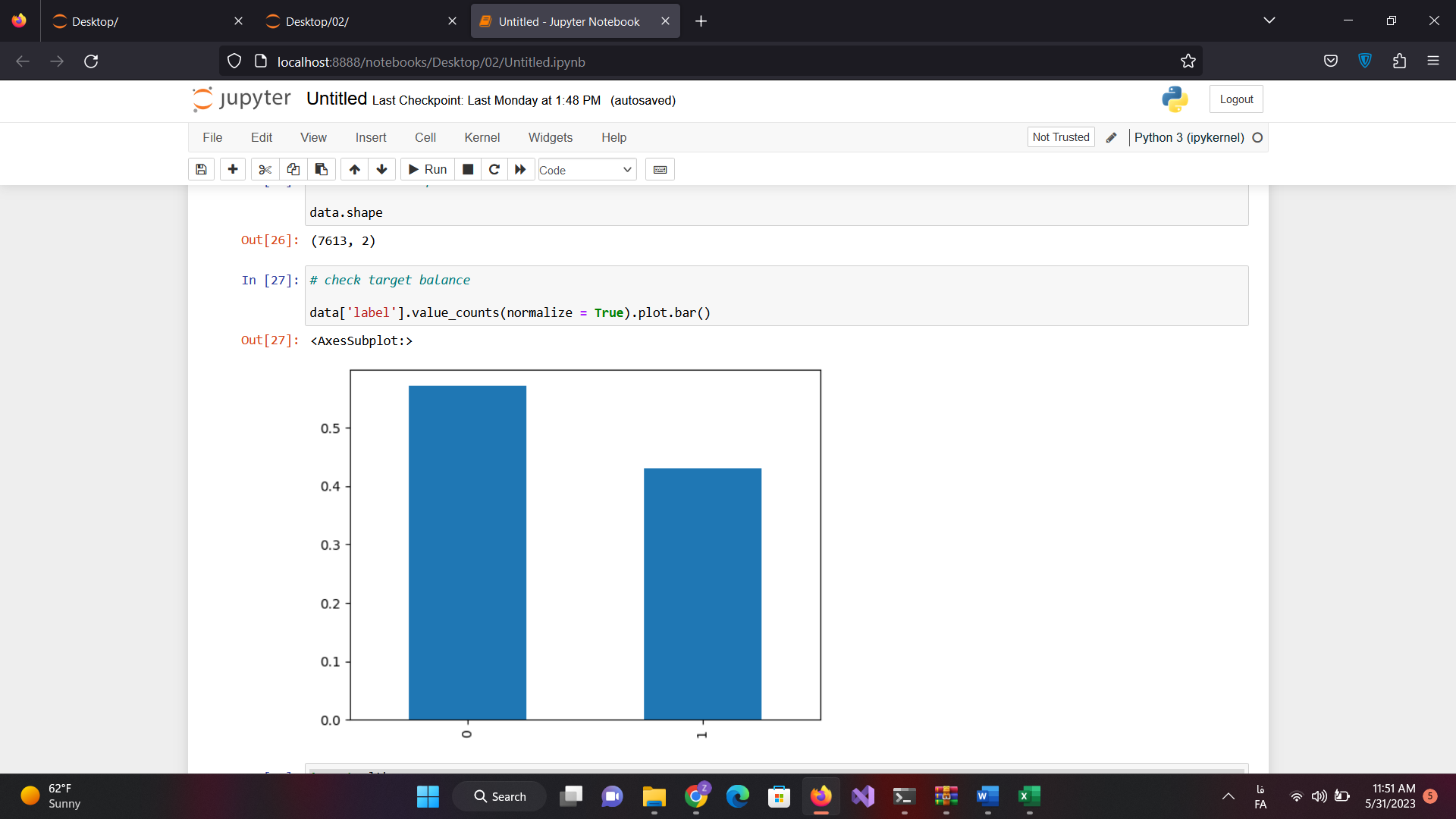
برای کلاس بندی ما یک داده داریم که شامل یکی متن هست و لیبل خورده که ایا هرکدوم از ان ها تارگت ما هست یا خیر و این داده ها را میگیریم و ستون های اضافی را پاک میکنیم.



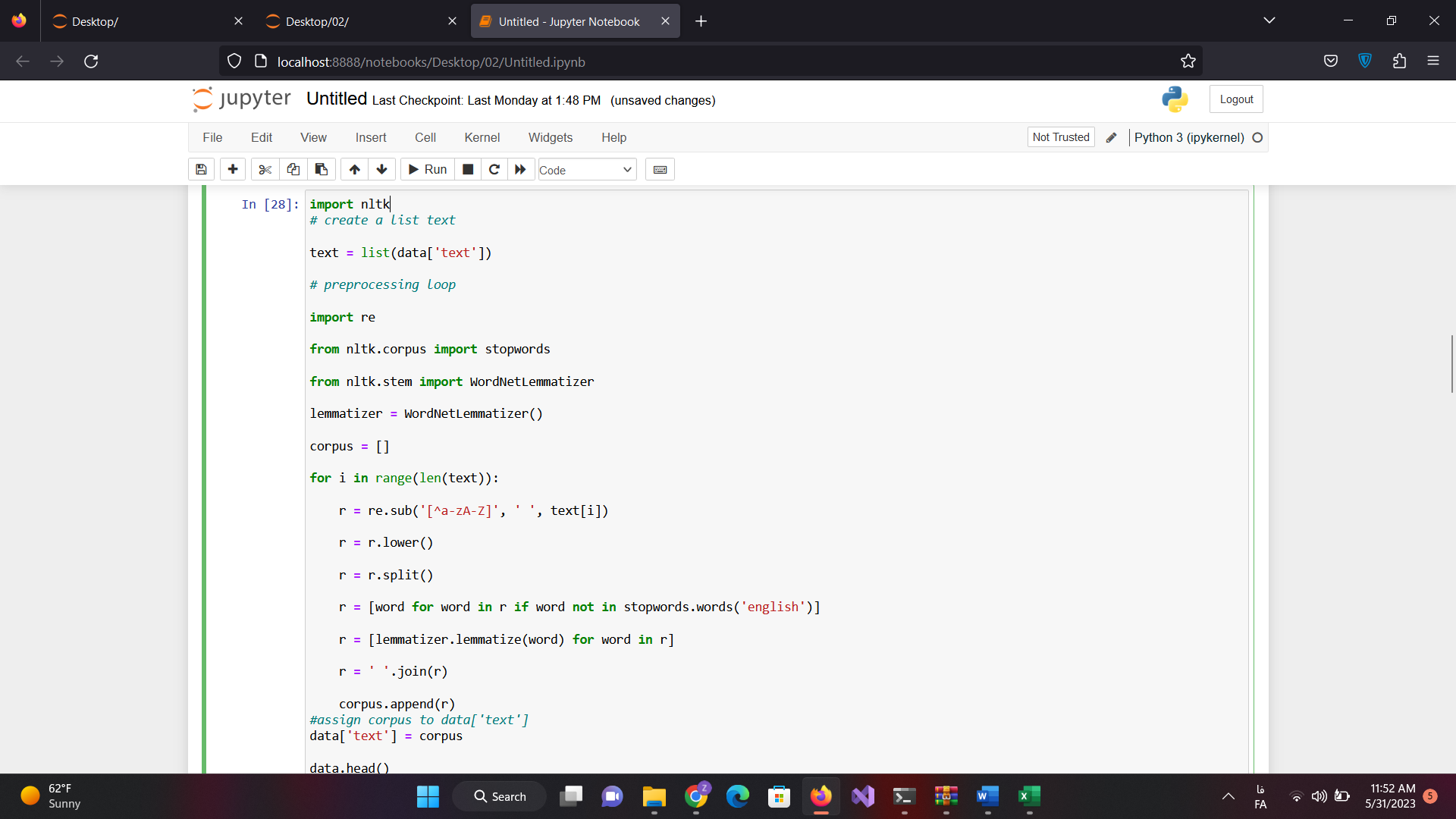
بیایید مقداری EDA اساسی انجام دهیم تا ببینیم آیا مقادیر گمشده در مجموعه داده وجود دارد یا خیر و تعادل هدف چیست.



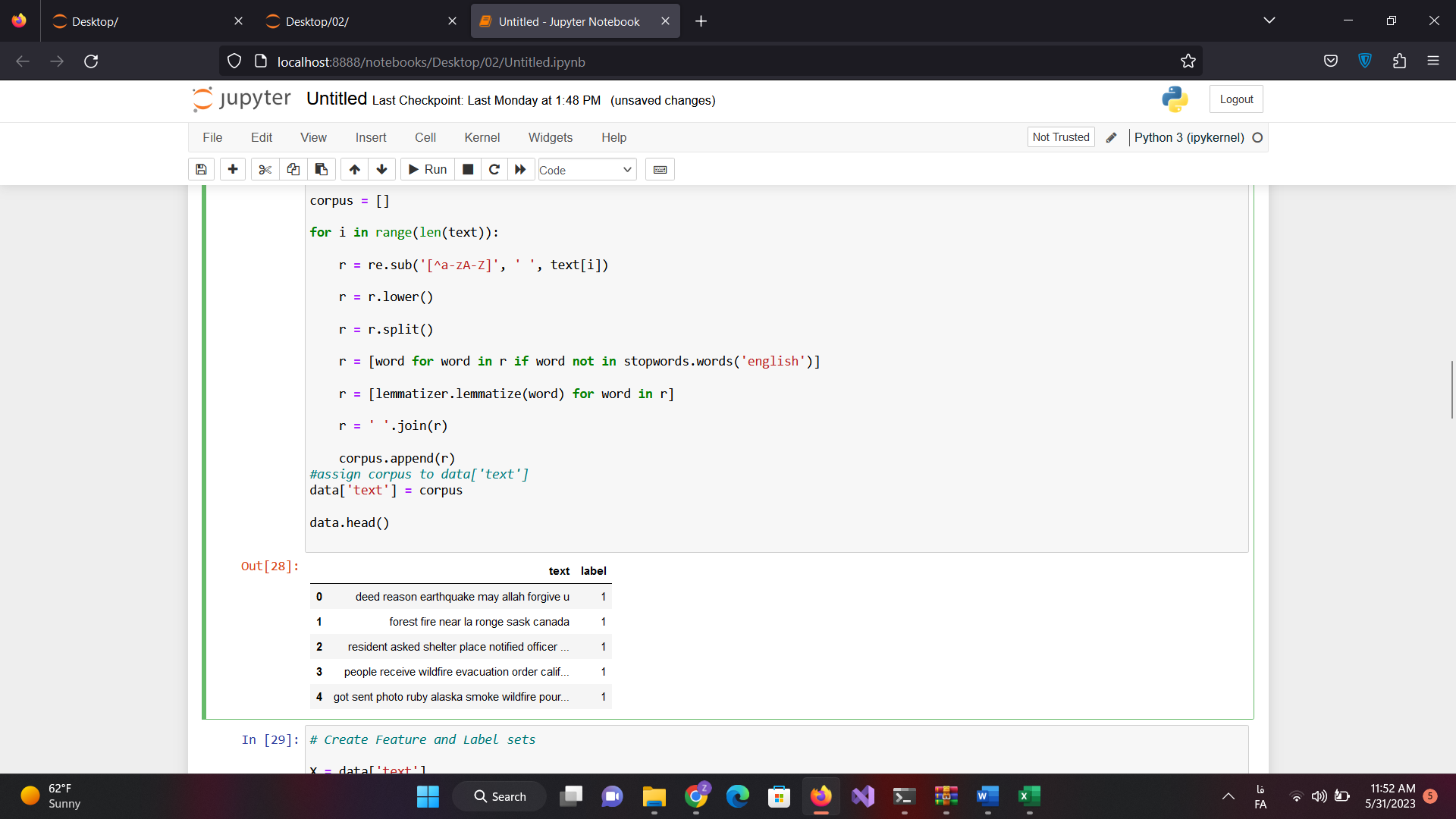
باکس بار تارگت را رسم میکنیم.



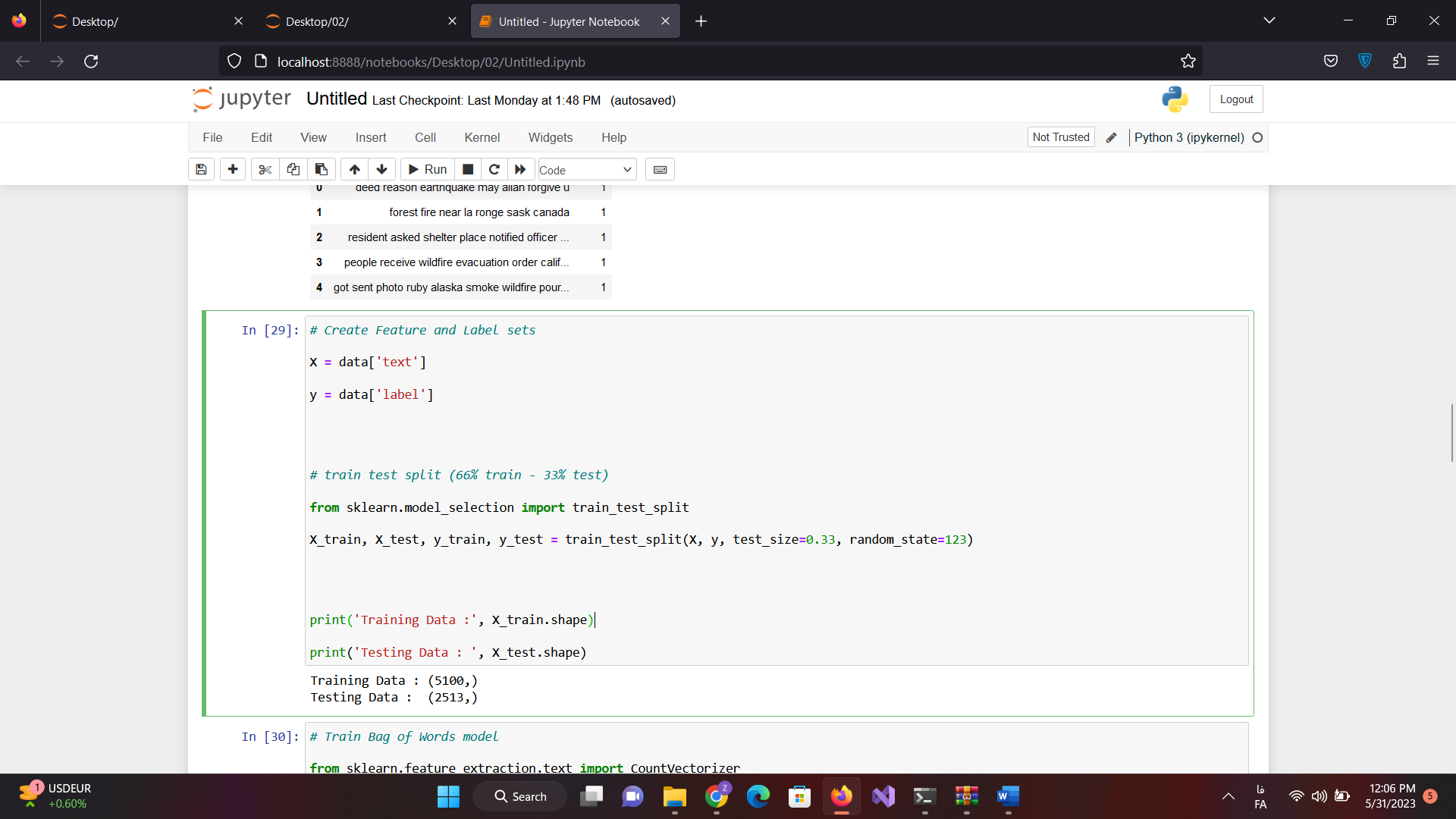
کد بعدی متن‌های موجود در ستون "text" یک دیتافریم را پیش‌پردازش می‌کند. ابتدا، عملیاتی صورت می‌گیرد تا عبارات غیر حروف الفبا را حذف و حروف را به حالت کوچک تبدیل کند. سپس، متن به کلمات تقسیم می‌شود و کلماتی که در لیست کلمات از کتابخانه "stopwords" (مثل "the"، "a"، "an" و...) وجود دارند حذف می‌شوند. سپس، کلمات به صورت lemmatizer (به ریشه تبدیل شده) می‌شوند. در نهایت، متن پیش‌پردازش شده در لیست "corpus" ذخیره می‌شود و ستون "text" دیتافریم با مقادیر پیش‌پردازش شده جدید به روزرسانی می‌شود.



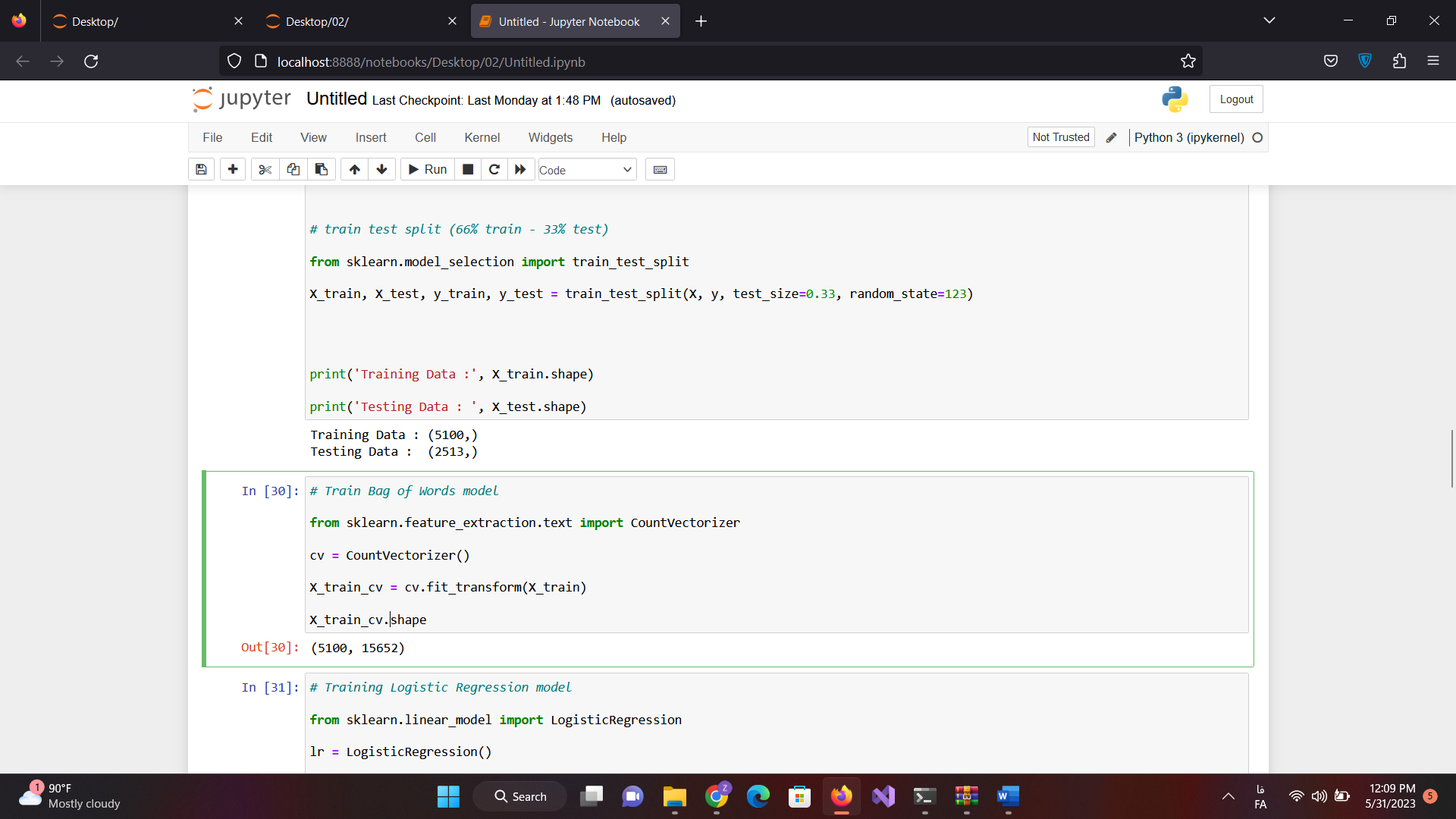
خروجی آن به صورت زیر است:



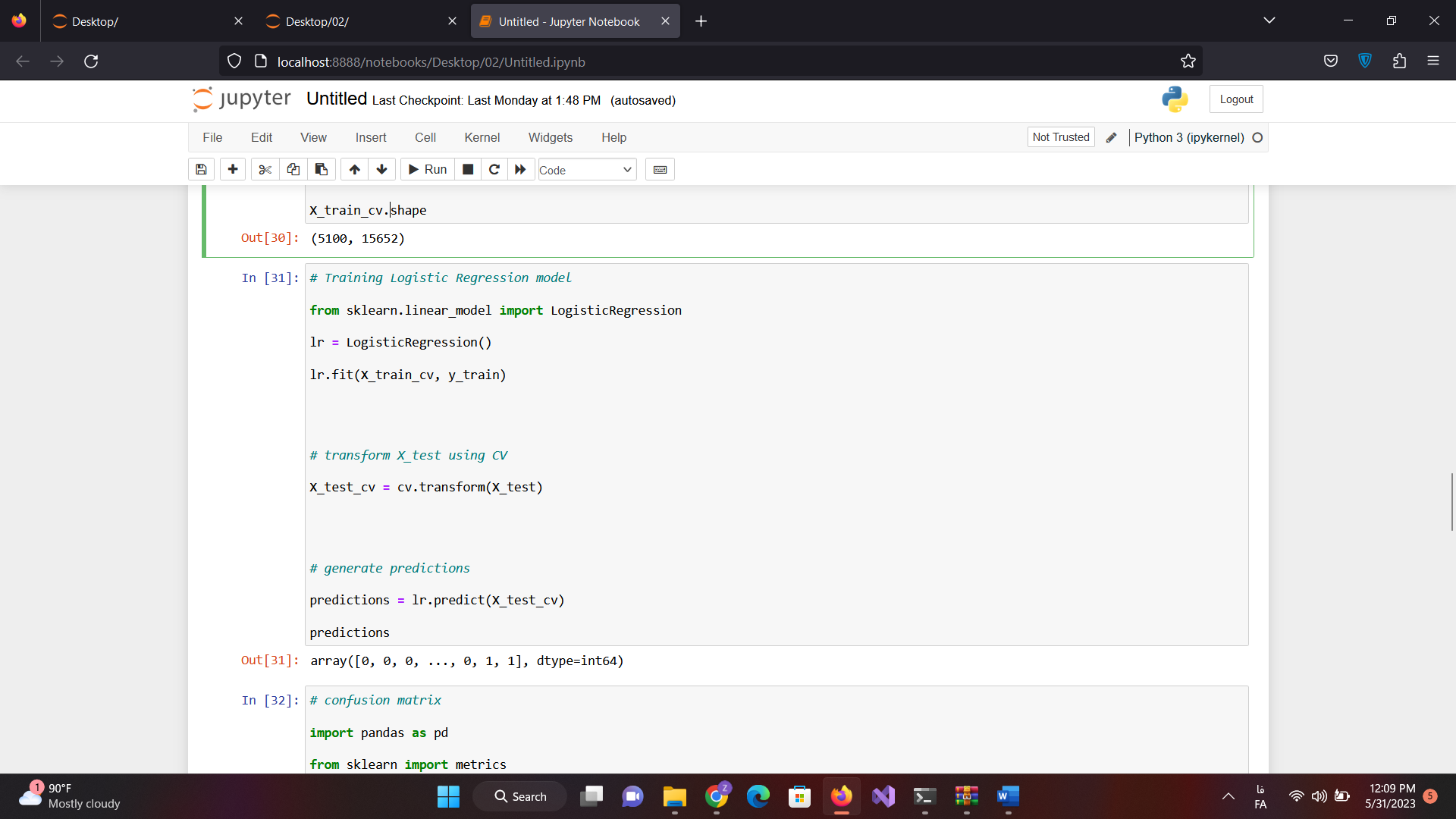
قسمت بعدی داده را به دو قسمت تقسیم میکنیم یکی برای تست و یکی برای ترین کردن داده ها. که تعداد داد های تست و ترین در خروجی مشخص شده است.



در این مثال، ما از روش Bag of Words (CountVectorizer) برای تبدیل متن خام به ویژگی‌های عددی استفاده می‌کنیم. این روش با توجه به تعداد تکرار کلمات در متن، برداری از ویژگی‌ها ایجاد می‌کند. این بردارها می‌توانند به عنوان ورودی برای مدل‌های یادگیری ماشین استفاده شوند.



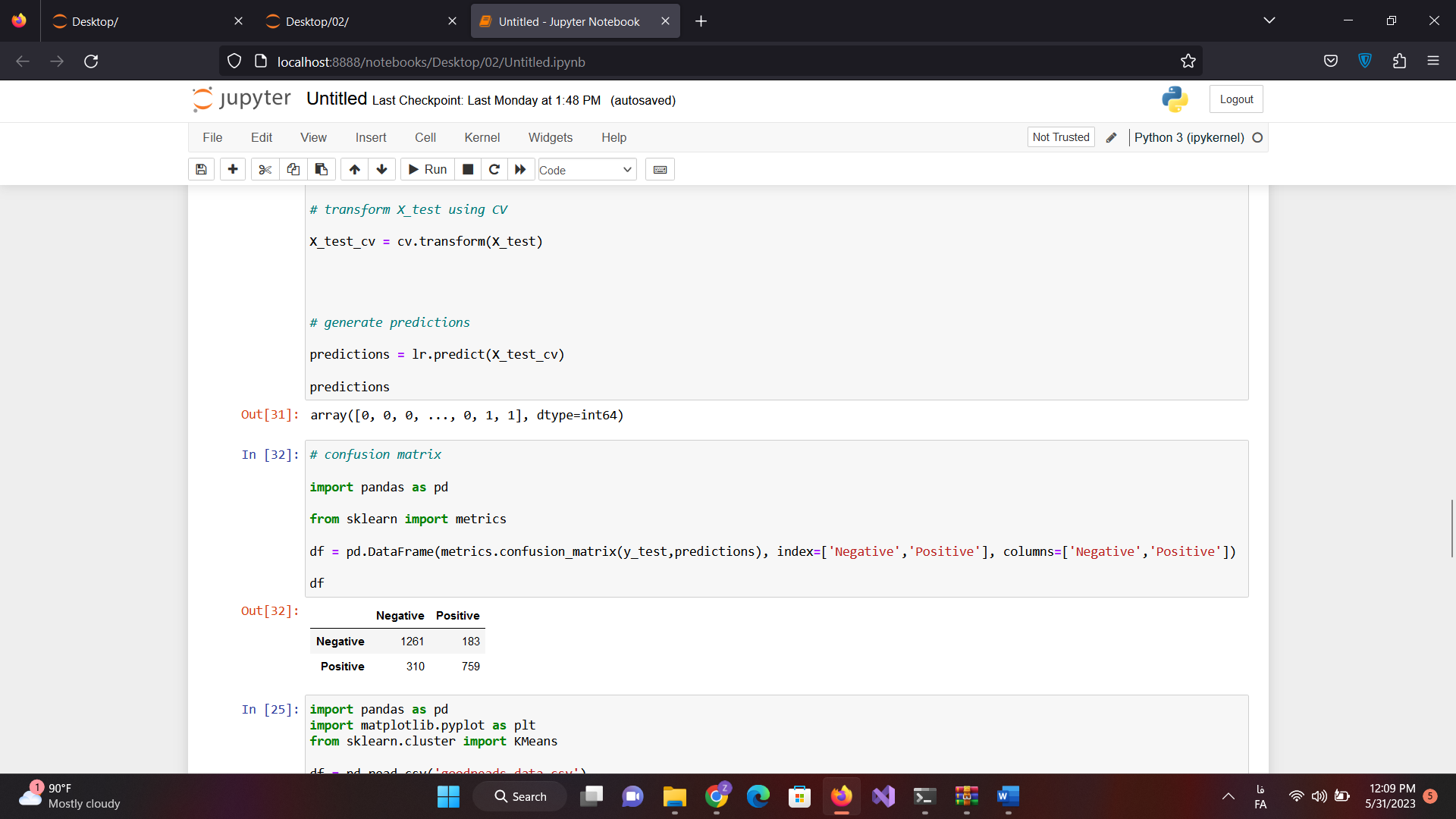
در این بخش یک مدل رگرسیون لجستیک را آموزش می دهیم و ماتریس سردرگمی مدل آموزش دیده را ارزیابی می کنیم. رگرسیون لجستیک یک الگوریتم طبقه بندی است که معمولاً در وظایف یادگیری ماشین استفاده می شود. ماتریس سردرگمی ارزیابی جامعی از عملکرد مدل را با نشان دادن تعداد مثبت های درست، منفی های درست، مثبت های کاذب و منفی های کاذب ارائه می دهد.



برای آموزش مدل رگرسیون لجستیک، از ویژگی های عددی استخراج شده از نمایش کیسه کلمات استفاده می کنیم. این ویژگی های عددی به عنوان ورودی مدل عمل می کنند. مدل از این ویژگی ها یاد می گیرد و بر اساس الگوهای آموخته شده در داده های آموزشی پیش بینی می کند.

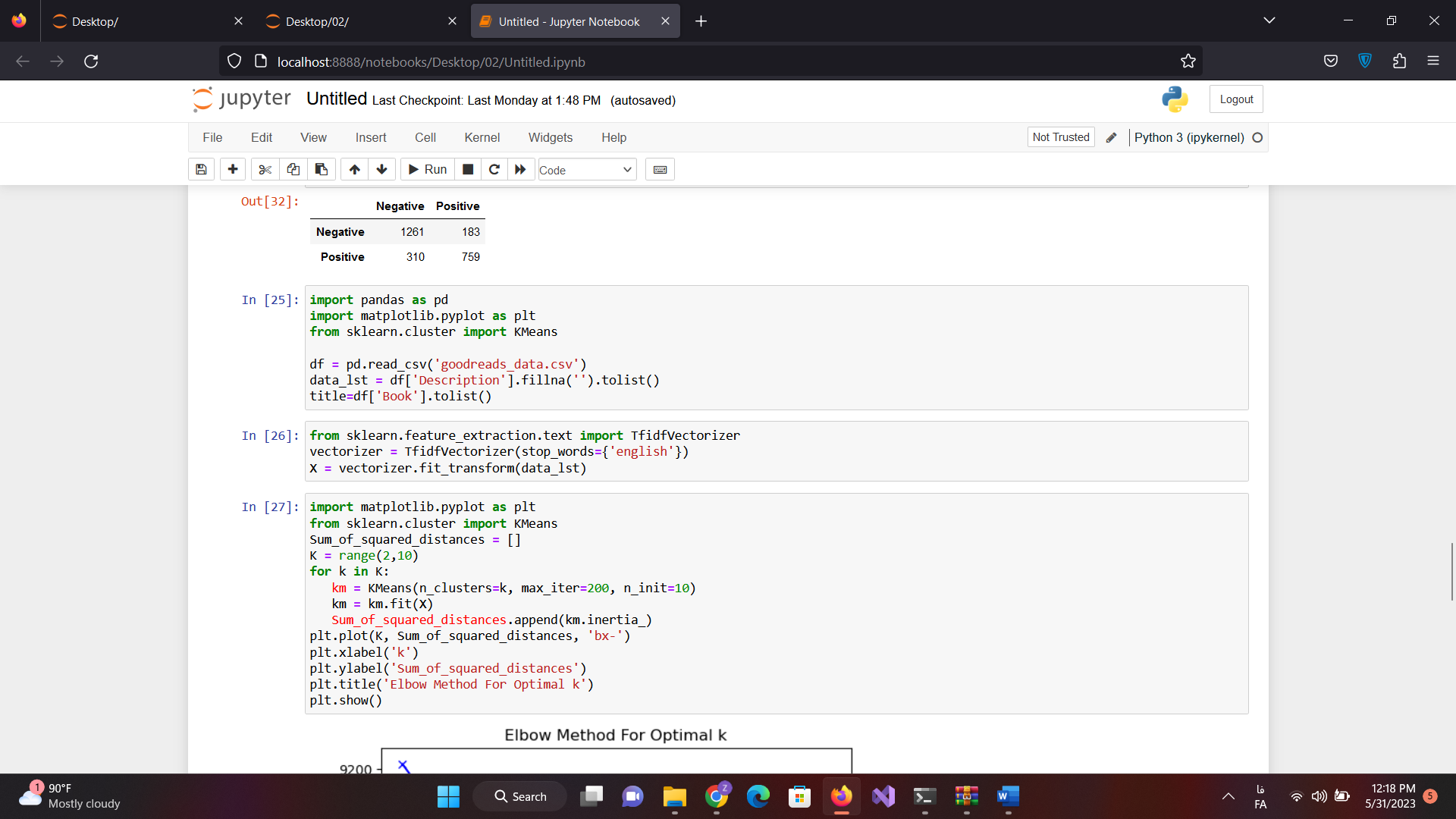
پس از آموزش مدل، عملکرد آن را با استفاده از ماتریس سردرگمی ارزیابی می کنیم. ماتریس سردرگمی جدولی است که برچسب های واقعی داده ها را با برچسب های پیش بینی شده تولید شده توسط مدل مقایسه می کند. این به ما کمک می کند تا دقت، دقت، فراخوانی و امتیاز F1 مدل را درک کنیم. ماتریس از چهار مقدار تشکیل شده است: مثبت واقعی (TP)، منفی واقعی (TN)، مثبت کاذب (FP) و منفی کاذب (FN).

با تجزیه و تحلیل ماتریس سردرگمی، می توانیم ارزیابی کنیم که مدل از نظر پیش بینی صحیح موارد مثبت و منفی چقدر خوب عمل می کند. این ارزیابی به ما کمک می کند تا نقاط قوت و ضعف مدل را درک کنیم و در مورد استقرار یا بهبودهای بیشتر آن تصمیمات آگاهانه بگیریم.

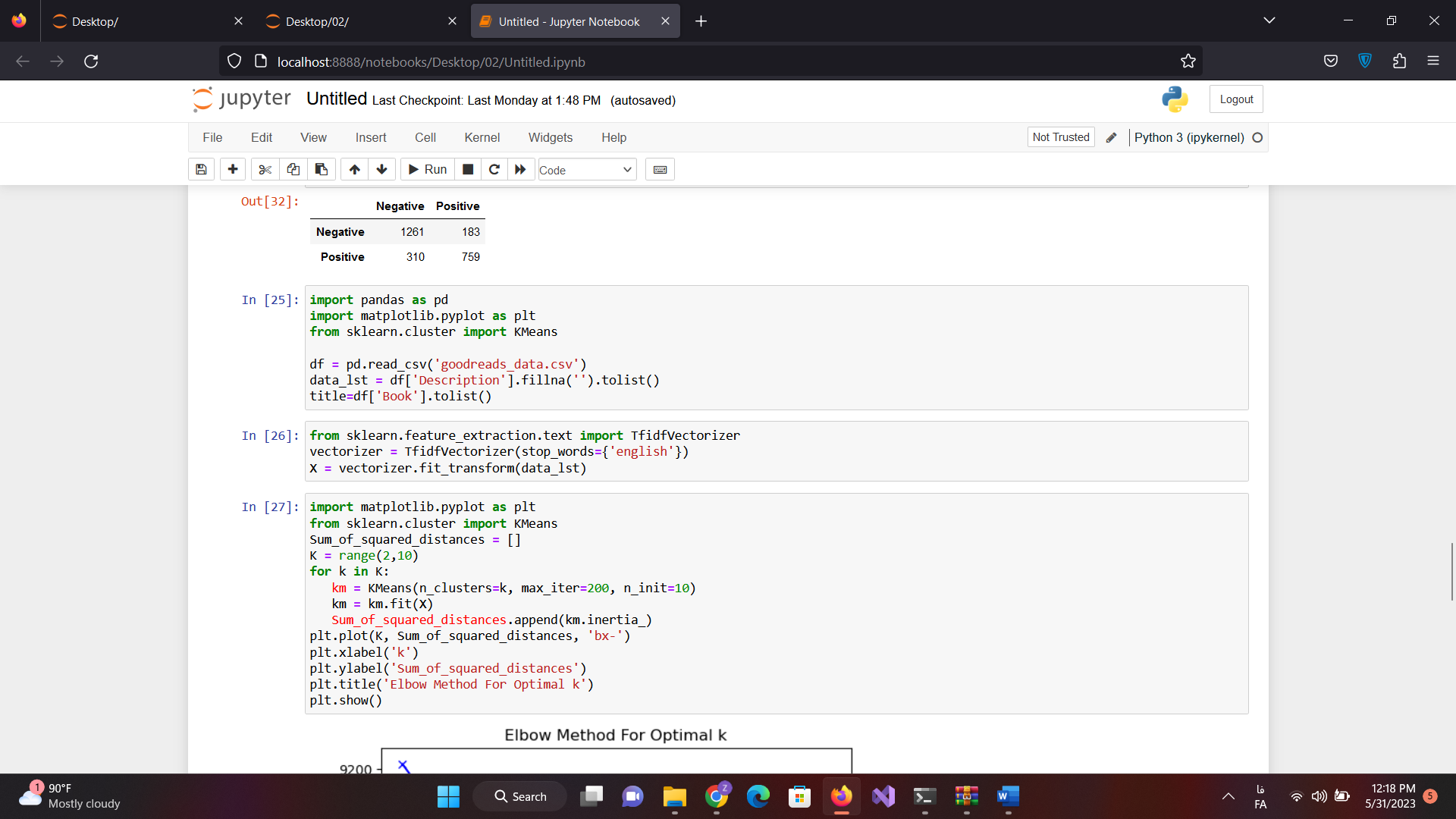


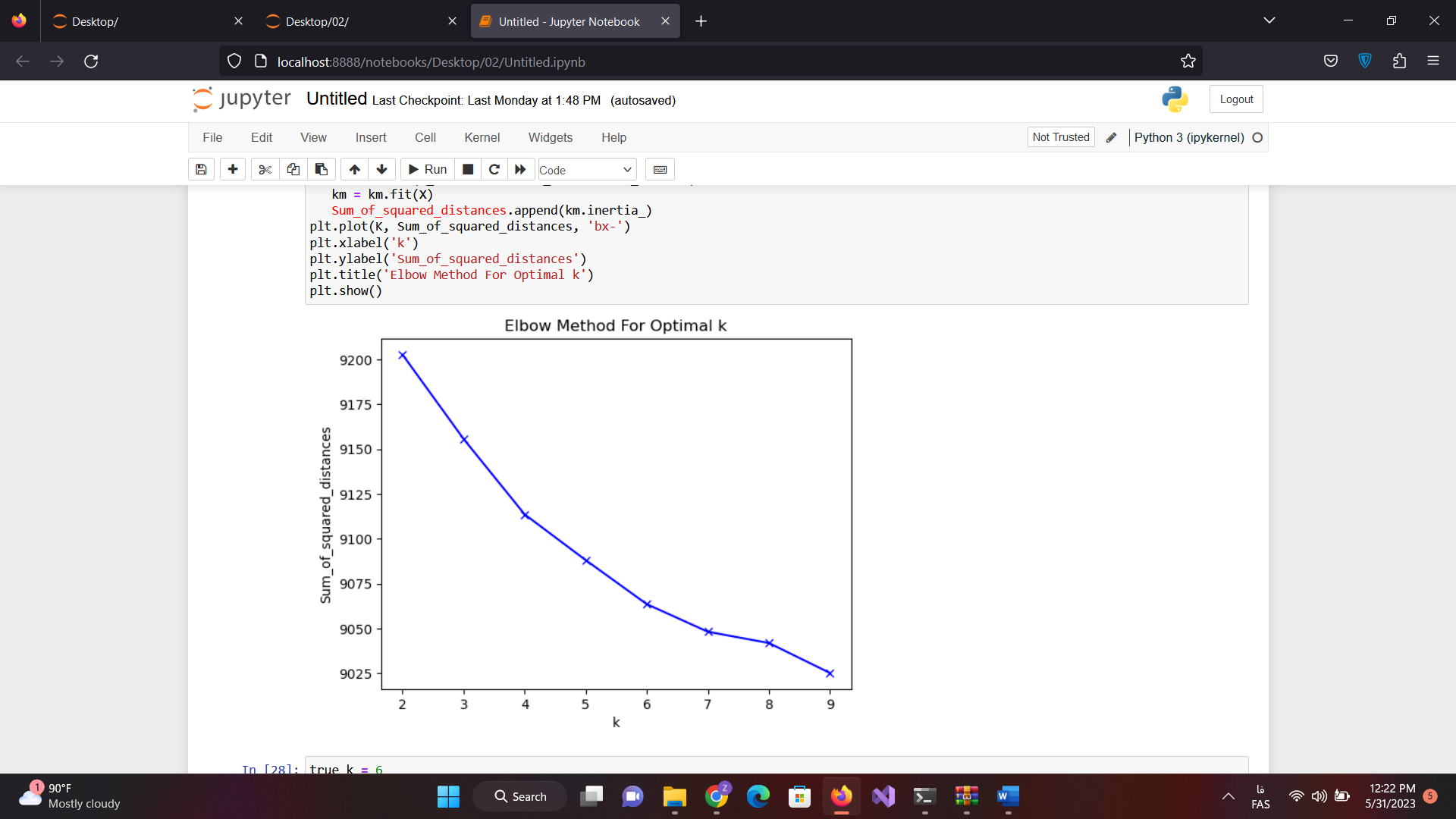
میریم سراغ قسمت بعدی یعنی خوشه بندی

که دیتاستی که میگیریم یک سری داده هست که شامل یک سری کتاب است که اسم کتاب و اسم نویسنده و سایتی که در آن کتاب هست و توضیحات کتاب است که ما در اینجا توضیحات کتاب را به عنوان متن و اسم کتاب را به عنوان تایتل میگیریم.

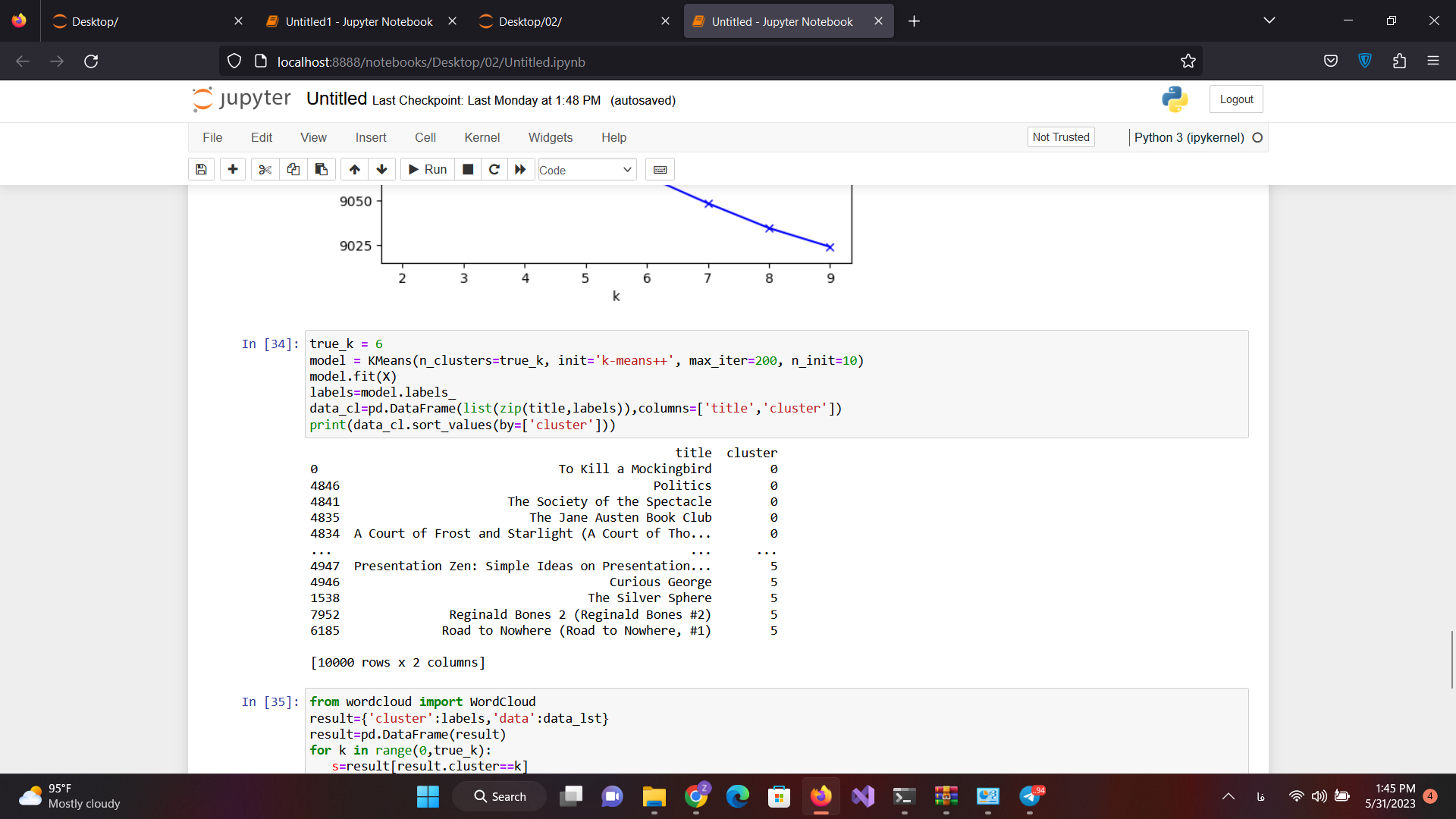


و بعد استاپ وردها را در می آوریم و روی بقیه داده ها tf-idf میزنیم. و بعد از آن با یکی از روش های k-means میایم ببنیم که چه تعداد خوشه بندی بکنیم بهتر است و نمودار آن را رسم میکنیم.





و برحسب نمودار 6 خوشه بندی برای این داده ها انجام میدهیم.



و بعد برای اینک ببنیم چه متن هایی در هر خوشه بندی با جزئیات بیشتر از کتابخونه WordCloud استفاده میکنیم.

