**تمرین سوم درس داده کاوی محمدباربد امیرمزلقانی – 810102348**

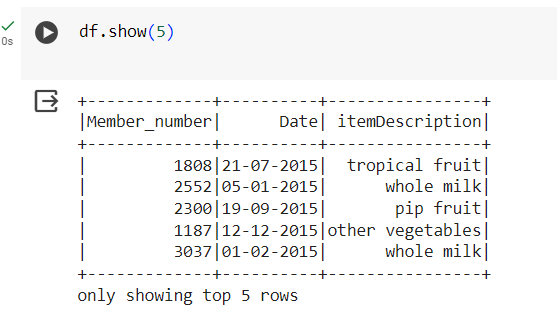
**بخش عملی**

**سوال اول)**

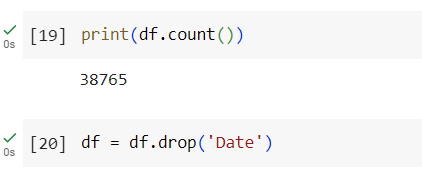
در گام اول برای پیاده سازی در spark لازم است آن را نصب کنیم که مطابق با تصویر زیر است:



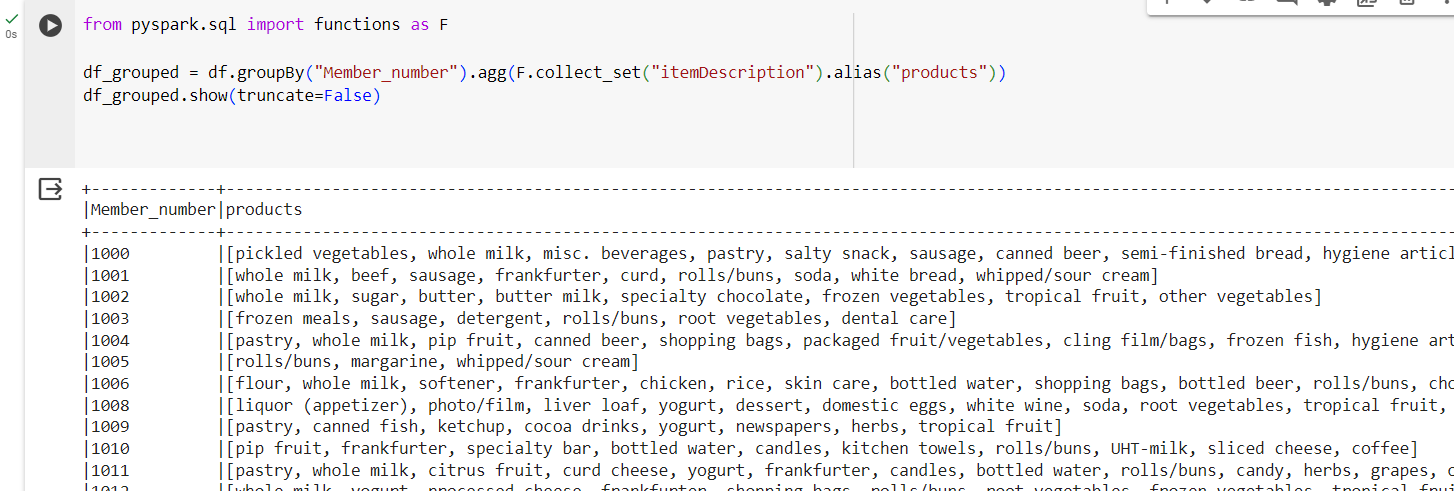
در بخش بعدی 5 سطر اول، تعداد سطرها و همچنین حذف ستون مربوط به تاریخ را انجام دهیم، ابتدا پنج سطر اول را با دستور زیر پرینت میکنیم :



در ادامه تعداد سطرها را خروجی میدهیم و ستون مربوط به تاریخ را حذف میکنیم:



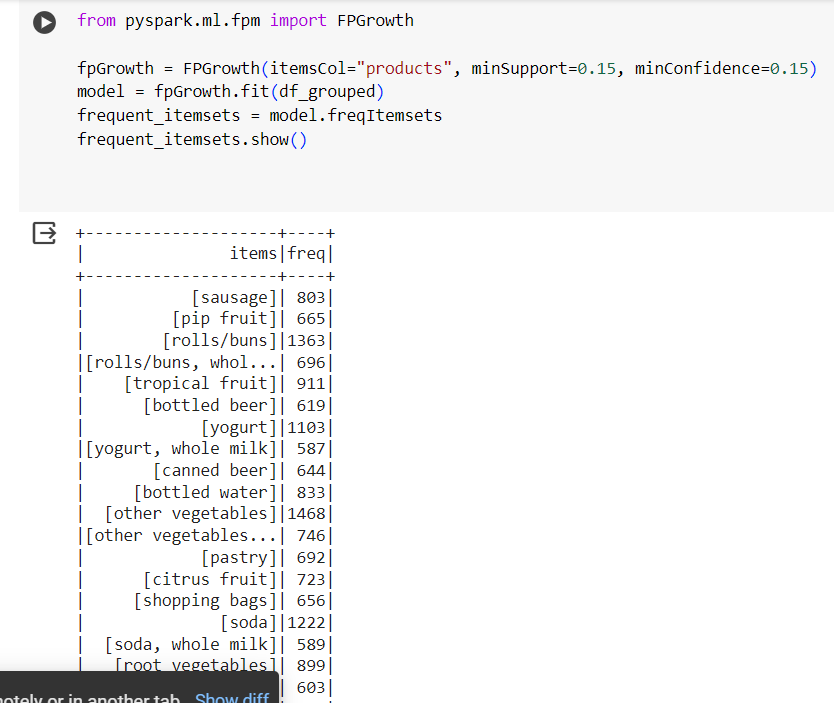
در ادامه به سراغ اگریگیت کردن روی آیتم لیست های مربوط به هر شماره کاربر میپردازیم ابتدا بر مبنای آی دی هر کاربر گروه میکنیم و در نهایت روی item هایی که خریده لیست ایجاد میکنیم(همچنین با توجه به fpgrowth که در ادامه استفاده میکنیم لازم است که ست باشد و تکراری نداشته باشد)



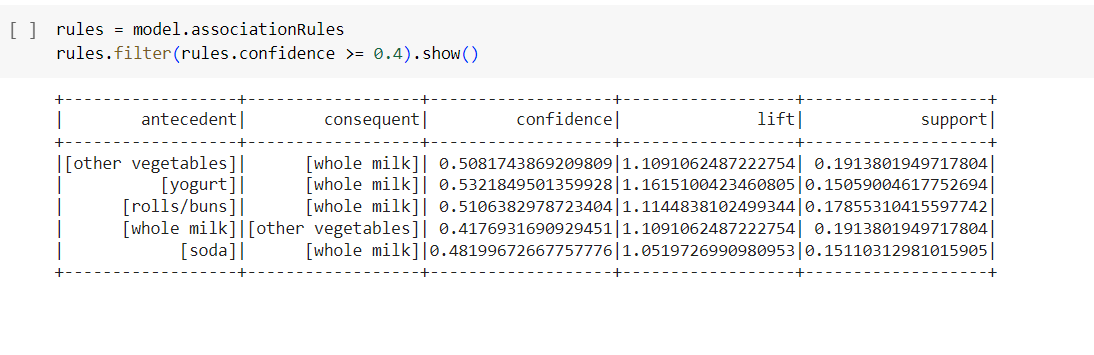
در مرحله بعد به سراغ ساخت دیتافریم جدید میرویم که ستون جدیدی هم دارد که product\_count است، سپس روی آن فیلتر میزنیم، که تنها آنهایی را برگرداند که تعداد آیتمهای آن از 10 بیشتر است :



در مرحله بعد به کمک فانکشن های built-in در اسپارک آن را فیت میکنیم و در نهایت نمایش میدهیم که به صورت زیر است :



در نهایت به سراغ قوانین میرویم که دوباره از ویژگی خود مدل اسپارک استفاده میکنیم و آن هایی که کانفیدنس بالای 0.4 دارند و آنها را نمایش میدهیم :



تمامی کدها در فایل نوتبوک موجود هستند.