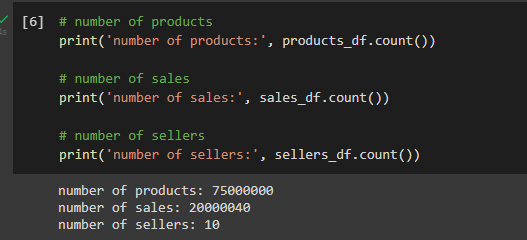
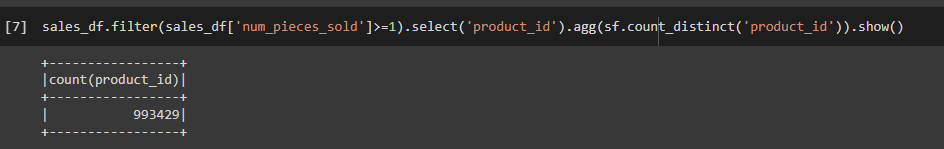
|  |
| --- |
| به‌نام خداوند بخشنده و مهربان  **تمرین اول:** مقدمه‌ای بر اسپارک **استاد:** محمد‌علی نعمت‌بخش  **درس:** پایگاه‌داده پیشرفته دستیاران: فاطمه‌ ابراهیمی، پریسا لطیفی، امیر سرتیپی  **نام و نام‌خانوادگی:** سید عمید اسدالهی مجد  **آدرس گیت:**<https://github.com/amidmajd/HW-2-spark-dataframes>  در ابتدا با انجام دستورات زیر یک Session جدید اسپارک می‌سازیم و داده‌های دانلودشده را خوانده و به DataFrame تبدیل می‌کنیم. سپس برای راحتی کار نوع ستون‌های قیمت، تعداد فروخته‌شده و مقدار فروش روزانه را به نوع داده‌ای Integer تبدیل می‌کنیم. |

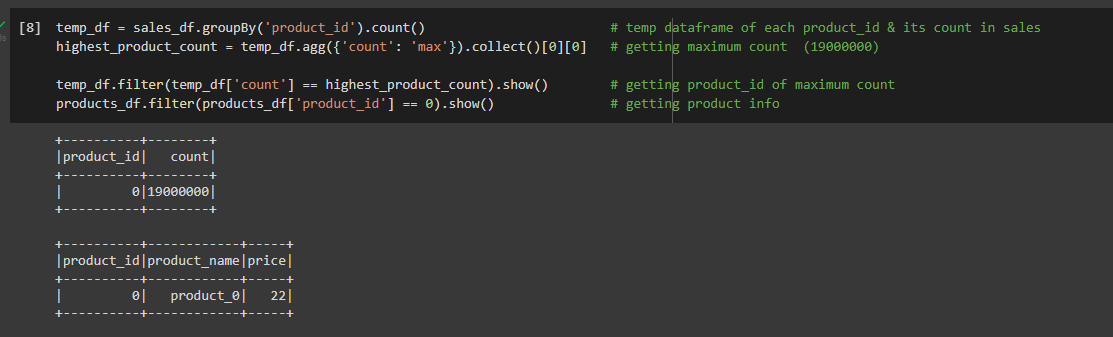
**۱. الف)**

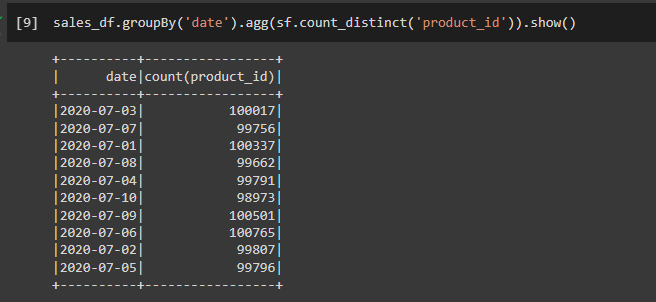


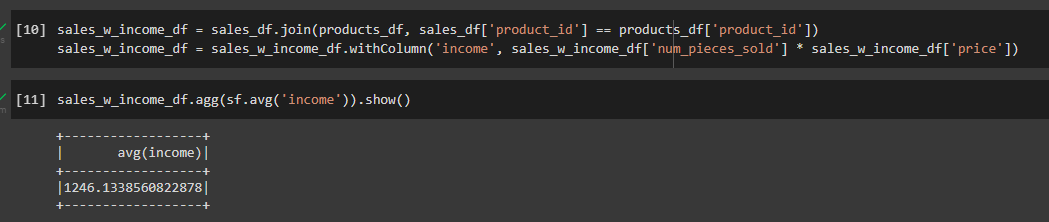
**۱. ب)**

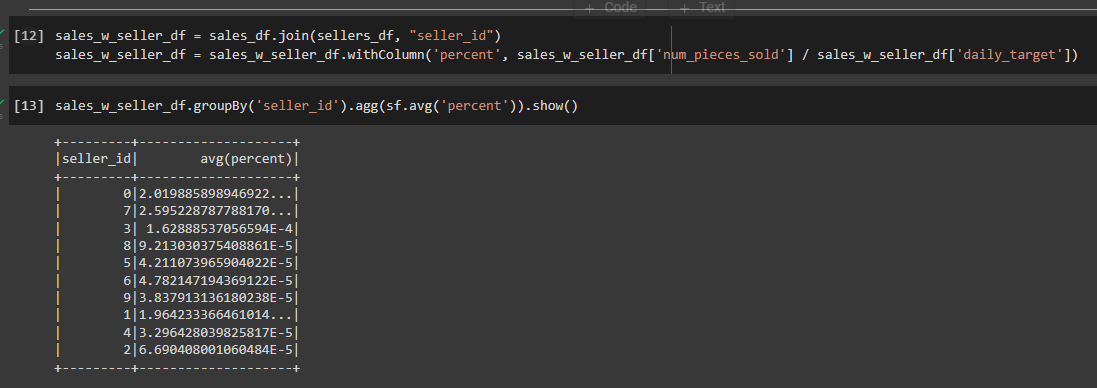


**۱. ج)** در این بخش ابتدا بر اساس product\_id گروه‌بندی می‌کنیم و یک دیتافریم موقت می‌سازیم و تعداد هر product\_id را با count() حساب می‌کنیم. سپس با استفاده از aggregate مقدار بیشینه تعداد محصولات موجود در داده‌های فروش را حساب می‌کنیم و در متغییر highest\_product\_count ذخیره می‌کنیم. سپس با دستور filter شماره product\_id مربوط به آن را از دیتافریم موقتی می‌یابیم و در انتها نام محصول را پیدا شد.

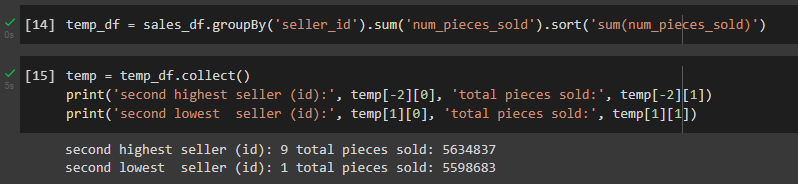


**۲.**

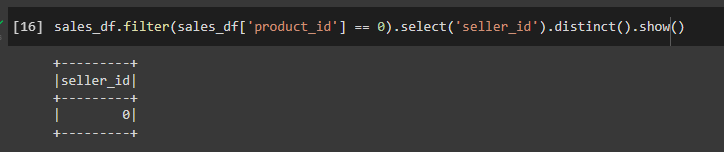
**۳.** در این بخش ابتدا دو دیتافریم فروش و محصول join زده شدند و سپس ستون جدید income با ضرب تعداد فروش هر محصول و قیمت آن به دیتا‌فریم جدید اضافه شد. در انتها میانگین درآمد سفارشات محاسبه شد.****

**۴. **

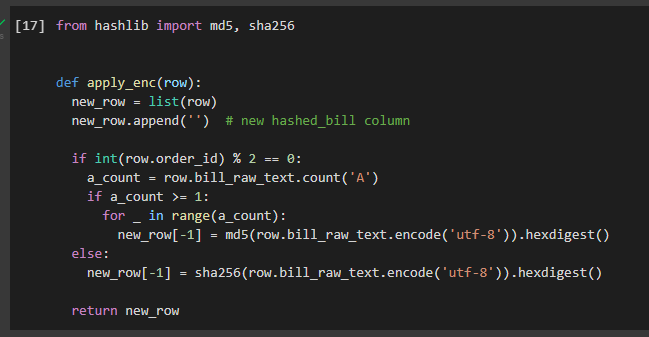
**۵. الف)** در این بخش ابتدا داده‌ها بر اساس seller\_id گروه‌بندی شدند و جمع تعداد فروش آن‌ها حساب شد. سپس بر اساس مجموع تعداد فروش مرتب شدند تا بتوان دومین پرفروش‌ترین و دومین کم‌فروش‌ترین فروشنده‌ها بدست آورده شود.



**۵. ب)**



**۶.** تابع انجام عملیات خواسته شده:



تبدیل به rdd و اعمال تابع با استفاده از map:



تبدیل مجدد خروجی به دیتافریم با نام ستون‌های صحیح:

