تمرین سوم داده کاوی – دکتر شاکری
محمد امین عرب خراسانی – ۸۱۰۱۰۲۲۰۵

تمرین عملی

الف) برای نمایش دیتاست، این دیتاست در goolge colab بارگزاری می‌شود. سپس به کمک دستور ()read.csv خوانده می‌شود. همچنین باید توجه داشت که به منظور استفاده از این کتابخانه ابتدا باید یک session اسپارک تعریف کرد.
برای نمایش ۵ سطر اول از ()df.show استفاده شده است. شکل زیر نتیجه‌ی حاصل را نشان می‌دهد.

```
+-------------+----------+----------------+
|Member_number|      Date| itemDescription|
+-------------+----------+----------------+
|         1808|21-07-2015|  tropical fruit|
|         2552|05-01-2015|      whole milk|
|         2300|19-09-2015|       pip fruit|
|         1187|12-12-2015|other vegetables|
|         3037|01-02-2015|      whole milk|
+-------------+----------+----------------+
only showing top 5 rows
```

برای نمایش تعداد سطرها از ()df.count استفاده شده است، همچنین تعداد ستون‌ها نیز مشخص شده‌اند.

```
# of rows: 38765
# of columns: 3
```

با دستور ()df.drop ین کار صورت می‌گیرد. نام ستون‌ها بعد از این انجام این قسمت در زیر آورده شده است.

```
['Member_number', 'itemDescription']
```

ب) برای داشتن دیتافریمی که خرید مربوط به هر member-number را در یک ردیف قرار دهد از ()df.groupBy و در ادامه از ()agg و ()collect-set استفاده شده است تا موارد تکراری حذف شود زیرا در ادامه‌ی سوال مطرح شده است که محصولات جدا مطرح هستند. همچنین اگر این مقادیر یونیک نباشند در الگوریتم FPgrowth مشکل ایجاد می‌شود. دیتافریم جدید شامل ۳۸۹۸ ردیف می‌باشد.

در ادامه با توجه به موارد اشاره شده، member-numberهایی که تعداد خرید یونیک بالای ۱۰ داشته باشند با کمک توابع filter() و size() انتخاب می‌شوند. تعداد ردیف‌ها در دیتافریم جدید به ۱۳۱۳ تا می‌رسد و به معنای آن است که ۱۳۱۳ نفر، خرید حداقل ۱۰ کالای یونیک را داشته‌اند.

ج) ابتدا تابع FPGrowth فراخوانی می‌شود سپس پارامترهای مرتبط آن مشخص می‌شوند. سپس روی ستون دیتافریم نهایی اعمال می شوند. طبق الگوریتم اجرا شده ۶۷ تا frequent items وجود دارد.

```
+--------------------+----+
|               items|freq|
+--------------------+----+
|              [pork]| 250|
|     [bottled water]| 419|
|[bottled water, o...| 239|
|[bottled water, s...| 202|
|[bottled water, w...| 276|
|        [newspapers]| 296|
| [frozen vegetables]| 231|
|      [citrus fruit]| 354|
|[citrus fruit, wh...| 226|
|       [white bread]| 198|
|            [butter]| 263|
|        [rolls/buns]| 650|
|[rolls/buns, othe...| 372|
|[rolls/buns, othe...| 238|
|[rolls/buns, whol...| 416|
|    [tropical fruit]| 456|
|[tropical fruit, ...| 216|
|[tropical fruit, ...| 230|
|[tropical fruit, ...| 202|
|[tropical fruit, ...| 272|
+--------------------+----+
only showing top 20 rows
```

د) با توجه به اجرای الگوریتم FPGrowth در مرحله‌ی قبل، association rules نیز با دستور associationRules حاصل می‌شود.

```
+--------------------+------------------+--------------------+-------------------+--------------------+
|          antecedent|        consequent|          confidence|               lift|             support|
+--------------------+------------------+--------------------+-------------------+--------------------+
|  [other vegetables]|      [rolls/buns]|  0.5081967213114754|1.0265573770491803| 0.2833206397562833|
|  [other vegetables]|          [yogurt]|0.42349726775956287|0.9876588145085364| 0.2361005331302361|
|  [other vegetables]|      [whole milk]|  0.6516393442622951|1.0101563860878318| 0.3632901751713633|
|  [other vegetables]|            [soda]|0.43579234972677594|0.9649162819414111|0.24295506473724296|
|      [bottled beer]|      [whole milk]|  0.6597633136094675| 1.022749977295432|0.16984006092916984|
|[whipped/sour cream]|      [whole milk]|  0.6496815286624203|1.0071214251874356|0.15536938309215537|
|[rolls/buns, othe...|      [whole milk]|  0.6397849462365591|0.9917799697858349|0.18126428027418126|
|             [yogurt]|      [rolls/buns]|  0.4920071047957371| 0.993854351687389|0.21096725057121096|
|             [yogurt]|[other vegetables]|  0.5506216696269982|0.9876588145085363| 0.2361005331302361|
|             [yogurt]|            [soda]|  0.4404973357015986|0.9753338984421568|0.18888042650418888|
|             [yogurt]|      [whole milk]|  0.6749555950266429|1.0463007039787275| 0.2894135567402894|
|        [rolls/buns]|[other vegetables]|  0.5723076923076923|1.0265573770491803| 0.2833206397562833|
|        [rolls/buns]|      [whole milk]|                0.64|0.9921133412042504|0.31683168316831684|
|        [rolls/buns]|          [yogurt]|0.42615384615384616| 0.993854351687389|0.21096725057121096|
|        [rolls/buns]|            [soda]|  0.4523076923076923|1.0014839797639123|0.22391469916222392|
|           [sausage]|      [rolls/buns]|  0.5196304849884527|1.0496535796766744|0.17136329017517135|
|           [sausage]|          [yogurt]|0.48036951501154734|1.1202933804798607|0.15841584158415842|
|           [sausage]|[other vegetables]|  0.5750577367205543|1.0314901752924697|0.18964204112718963|
|           [sausage]|            [soda]|0.46882217090069284| 1.038049764574384| 0.1546077684691546|
|           [sausage]|      [whole milk]|  0.6374133949191686| 0.988103645252501| 0.2102056359482102|
+--------------------+------------------+--------------------+-------------------+--------------------+
only showing top 20 rows

53
```