

Instacart Market Basket Analysis

CS-6301 Practical Aspects of Data Science
Summer 2020

Assignment 4

Rajarshi Chattopadhyay
RXC170010

1. Frequent itemsets for products in orders dataset

The support threshold is kept 0.01 to handle the small probability of a frequent itemset.
The itemset min size is kept 2 as an item set should involve at least 2 items.

Apriori

Parameter specification:

confidence	minval	smax	arem	aval	originalSupport	maxtime	support	minlen	maxlen	target	ext
NA	0.1	1	none	FALSE	TRUE	5	0.01	2	10	frequent itemsets	TRUE

Frequent product itemsets ordered based on support: -

	items	support	transIdenticalToItemsets	count
[1]	{Bag of Organic Bananas,Organic Hass Avocado}	0.01939143	0.0009617795	62341
[2]	{Bag of Organic Bananas,Organic Strawberries}	0.01916965	0.0011997360	61628
[3]	{Banana,Organic Strawberries}	0.01746756	0.0010544737	56156
[4]	{Banana,Organic Avocado}	0.01660874	0.0009916407	53395
[5]	{Banana,Organic Baby Spinach}	0.01598663	0.0006501032	51395
[6]	{Bag of Organic Bananas,Organic Baby Spinach}	0.01566842	0.0009474710	50372
[7]	{Banana,Strawberries}	0.01282539	0.0015779779	41232
[8]	{Banana,Large Lemon}	0.01271589	0.0006581906	40880
[9]	{Organic Hass Avocado,Organic Strawberries}	0.01268914	0.0003586455	40794
[10]	{Bag of Organic Bananas,Organic Raspberries}	0.01259863	0.0006413937	40503
[11]	{Organic Baby Spinach,Organic Strawberries}	0.01186174	0.0003894398	38134
[12]	{Organic Baby Spinach,Organic Hass Avocado}	0.01085610	0.0002703061	34901
[13]	{Banana,Organic Fuji Apple}	0.01055811	0.0012258645	33943
[14]	{Organic Raspberries,Organic Strawberries}	0.01053323	0.0004718692	33863

- With a support threshold of 0.01 there are 14 baskets.
- People buy bananas often. The top 8 pairs with highest support contain bananas.
- All of the items are either fruits or vegetables.

2. Association rules for products in orders dataset

A low support threshold 0.0001 is used for less frequent items with a high confidence 0.6 to generate strong rules.

Mining returned only patterns up to length of 3 as time limit was reached.

set of 118 rules

rule length distribution (lhs + rhs):sizes
2 3
4 114

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.000	3.000	3.000	2.966	3.000	3.000

summary of quality measures:

support	confidence	coverage	lift	count
Min. :0.0001002	Min. :0.6005	Min. :0.0001521	Min. : 4.086	Min. : 322.0
1st Qu.:0.0001214	1st Qu.:0.6084	1st Qu.:0.0001911	1st Qu.: 64.864	1st Qu.: 390.2
Median :0.0001518	Median :0.6278	Median :0.0002338	Median : 170.099	Median : 488.0
Mean :0.0002091	Mean :0.6385	Mean :0.0003293	Mean : 330.143	Mean : 672.2
3rd Qu.:0.0002155	3rd Qu.:0.6586	3rd Qu.:0.0003258	3rd Qu.: 477.681	3rd Qu.: 692.8
Max. :0.0008504	Max. :0.7795	Max. :0.0013120	Max. :2167.152	Max. :2734.0

mining info:

data	ntransactions	support	confidence
transactions	3214874	1e-04	0.6

Ordered by Confidence

lhs	rhs	support	confidence	coverage	lift	count
[1] {Oh My Yog! Madagascar Vanilla Trilayer Yogurt, Oh My Yog! Organic Wild Quebec Blueberry Cream Top Yogurt & Fruit}	=> {Oh My Yog! Pacific Coast Strawberry Trilayer Yogurt}	0.0002385786	0.7794715	0.0003060773	877.10984	767
[2] {Organic Apples, Carrots and Parsnips Puree, Stage 1 Apples & Strawberries Organic Pureed Baby Food}	=> {Stage 1 Apples Sweet Potatoes Pumpkin & Blueberries Organic Pureed Baby Food}	0.0001377970	0.7559727	0.0001822778	435.46980	443
[3] {Oh My Yog! Gingered Pear Trilayer Yogurt, Oh My Yog! Organic Wild Quebec Blueberry Cream Top Yogurt & Fruit}	=> {Oh My Yog! Pacific Coast Strawberry Trilayer Yogurt}	0.0001838330	0.7509530	0.0002447996	845.01898	591
[4] {All Natural Apricot Sparkling Water, Sparkling Water Berry}	=> {Sparkling Water Grapefruit}	0.0001359307	0.7344538	0.0001850772	31.11478	437
[5] {Oh My Yog! Gingered Pear Trilayer Yogurt, Oh My Yog! Pacific Coast Strawberry Trilayer Yogurt}	=> {Oh My Yog! Organic Wild Quebec Blueberry Cream Top Yogurt & Fruit}	0.0001838330	0.7224939	0.0002544423	1022.32694	591
[6] {Grapefruit Sparkling Water, Natural Lemon Flavored Sparkling Water}	=> {Lemon Sparkling Water}	0.0001418407	0.7158556	0.0001981415	242.35315	456
[7] {Organic 4 Months Butternut Squash Carrots Apples + Prunes Baby Food, Organic Apples, Carrots and Parsnips Puree}	=> {Stage 1 Apples Sweet Potatoes Pumpkin & Blueberries Organic Pureed Baby Food}	0.0001875657	0.7144550	0.0002625297	411.55397	603
[8] {Blackberry Yogurt, Total 2% All Natural Greek Strained Yogurt with Honey}	=> {Total 2% with Strawberry Lowfat Greek Strained Yogurt}	0.0001545939	0.7009873	0.0002205374	75.38841	497
[9] {Unsweet Peach Water, Unsweetened Strawberry Kiwi Water}	=> {Unsweetened Watermelon Water}	0.0001182006	0.6959707	0.0001698356	540.18785	380
[10] {Oh My Yog! Gingered Pear Trilayer Yogurt, Oh My Yog! Madagascar Vanilla Trilayer Yogurt}	=> {Oh My Yog! Pacific Coast Strawberry Trilayer Yogurt}	0.0001567713	0.6942149	0.0002258253	781.17373	584

Ordered by Lift

lhs	rhs	support	confidence	coverage	lift	count
[1] {Organic Cashew Nondairy Blueberry Yogurt}	=> {Organic Nondairy Strawberry Cashew Yogurt}	0.0001188227	0.6161290	0.0001928536	2167.1523	382
[2] {Organic Cottage Cheese Blueberry Acai Chia}	=> {Organic Strawberry Chia Lowfat 2% Cottage Cheese}	0.0001608150	0.6162098	0.0002609745	1696.0931	517
[3] {Grain Free Turkey Canned Cat Food}	=> {Grain Free Chicken Formula Cat Food}	0.0001007816	0.6428571	0.0001567713	1132.4409	324
[4] {Unsweetened Whole Milk Mixed Berry Greek Yogurt, Unsweetened Whole Milk Strawberry Yogurt}	=> {Unsweetened Whole Milk Peach Greek Yogurt}	0.0001483728	0.6489796	0.000226248	1085.5295	477
[5] {Unsweetened Whole Milk Blueberry Greek Yogurt, Unsweetened Whole Milk Peach Greek Yogurt}	=> {Unsweetened Whole Milk Strawberry Yogurt}	0.0001517944	0.6404199	0.0002370233	1047.7707	488
[6] {Unsweetened Whole Milk Blueberry Greek Yogurt, Unsweetened Whole Milk Mixed Berry Greek Yogurt}	=> {Unsweetened Whole Milk Peach Greek Yogurt}	0.0001402854	0.6246537	0.0002245811	1044.8403	451
[7] {Unsweetened Whole Milk Blueberry Greek Yogurt, Unsweetened Whole Milk Strawberry Yogurt}	=> {Unsweetened Whole Milk Peach Greek Yogurt}	0.0001517944	0.6177215	0.0002457328	1033.2450	488
[8] {Oh My Yog! Gingered Pear Trilayer Yogurt, Oh My Yog! Pacific Coast Strawberry Trilayer Yogurt}	=> {Oh My Yog! Organic Wild Quebec Blueberry Cream Top Yogurt & Fruit}	0.0001838330	0.7224939	0.0002544423	1022.3269	591
[9] {Unsweetened Whole Milk Blueberry Greek Yogurt, Unsweetened Whole Milk Mixed Berry Greek Yogurt}	=> {Unsweetened Whole Milk Strawberry Yogurt}	0.0001365528	0.6080332	0.0002245811	994.7838	439
[10] {Mighty 4 Essential Tots Spinach, Kiwi, Barley & Greek Yogurt Nutrition Blend, Mighty 4 Sweet Potato, Blueberry, Millet & Greek Yogurt Tots Snack}	=> {Mighty 4 Kale, Strawberry, Amaranth & Greek Yogurt Tots Snack}	0.0001290875	0.6250000	0.0002065400	910.8324	415

The rules with higher lift indicate a strong association between the items.

A higher support threshold 0.001 is used for more frequent items with a lower confidence 0.4 to generate weaker rules.

set of 39 rules

rule length distribution (lhs + rhs):sizes

2 3 4
7 30 2

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.000	3.000	3.000	2.872	3.000	4.000

summary of quality measures:

support	confidence	coverage	lift	count
Min. :0.001006	Min. :0.4001	Min. :0.001954	Min. : 2.779	Min. : 3235
1st Qu.:0.001161	1st Qu.:0.4159	1st Qu.:0.002534	1st Qu.: 3.057	1st Qu.: 3733
Median :0.001267	Median :0.4411	Median :0.002890	Median : 3.683	Median : 4073
Mean :0.001465	Mean :0.4438	Mean :0.003331	Mean :18.429	Mean : 4708
3rd Qu.:0.001522	3rd Qu.:0.4581	3rd Qu.:0.003517	3rd Qu.:27.591	3rd Qu.: 4892
Max. :0.003549	Max. :0.5942	Max. :0.008023	Max. :75.648	Max. :11409

mining info:

data n	transactions	support	confidence
transactions	3214874	0.001	0.4

Ordered by Confidence

	lhs	rhs	support	confidence	coverage	lift	count
[1]	{Total 2% Lowfat Greek Strained Yogurt With Blueberry, Total 2% Lowfat Greek Strained Yogurt with Peach}	=> {Total 2% with Strawberry Lowfat Greek Strained Yogurt}	0.001161165	0.5942375	0.001954042	63.907895	3733
[2]	{Lime Sparkling Water, Sparkling Lemon Water}	=> {Sparkling Water Grapefruit}	0.001379214	0.5188392	0.002658269	21.980375	4434
[3]	{Honeycrisp Apple, Strawberries}	=> {Banana}	0.001149967	0.4905786	0.002344104	3.337421	3697
[4]	{Organic Fuji Apple, Strawberries}	=> {Banana}	0.001141258	0.4901804	0.002328241	3.334712	3669
[5]	{Organic Hass Avocado, Organic Raspberries, Organic Strawberries}	=> {Bag of Organic Bananas}	0.001190715	0.4850482	0.002454840	4.109550	3828
[6]	{Sparkling Lemon Water, Sparkling Water Grapefruit}	=> {Lime Sparkling Water}	0.001379214	0.4807025	0.002869164	33.201521	4434
[7]	{Total 2% Lowfat Greek Strained Yogurt with Peach, Total 2% with Strawberry Lowfat Greek Strained Yogurt}	=> {Total 2% Lowfat Greek Strained Yogurt With Blueberry}	0.001161165	0.4658098	0.002492788	72.141821	3733
[8]	{Zero Calorie Cola}	=> {Soda}	0.001234885	0.4638934	0.002662002	41.668546	3970
[9]	{Organic Hass Avocado, Organic Navel Orange}	=> {Bag of Organic Bananas}	0.001175474	0.4635672	0.002535714	3.927554	3779
[10]	{Cucumber Kirby, Organic Fuji Apple}	=> {Banana}	0.001024924	0.4625860	0.002215639	3.146986	3295

Ordered by Lift

	lhs	rhs	support	confidence	coverage	lift	count
[1]	{Non Fat Acai & Mixed Berries Yogurt}	=> {Icelandic Style Skyr Blueberry Non-fat Yogurt}	0.001220888	0.4531286	0.002694351	75.64789	3925
[2]	{Non Fat Raspberry Yogurt}	=> {Icelandic Style Skyr Blueberry Non-fat Yogurt}	0.002247055	0.4411064	0.005094134	73.64084	7224
[3]	{Total 2% Lowfat Greek Strained Yogurt with Peach, Total 2% with Strawberry Lowfat Greek Strained Yogurt}	=> {Total 2% Lowfat Greek Strained Yogurt With Blueberry}	0.001161165	0.4658098	0.002492788	72.14182	3733
[4]	{Nonfat Icelandic Style Strawberry Yogurt}	=> {Icelandic Style Skyr Blueberry Non-fat Yogurt}	0.001418096	0.4275532	0.003316771	71.37819	4559
[5]	{Total 2% Lowfat Greek Strained Yogurt With Blueberry, Total 2% with Strawberry Lowfat Greek Strained Yogurt}	=> {Total 2% Lowfat Greek Strained Yogurt with Peach}	0.001161165	0.4000643	0.002902447	64.60500	3733
[6]	{Total 2% Lowfat Greek Strained Yogurt With Blueberry, Total 2% Lowfat Greek Strained Yogurt with Peach}	=> {Total 2% with Strawberry Lowfat Greek Strained Yogurt}	0.001161165	0.5942375	0.001954042	63.90789	3733
[7]	{Total 2% Lowfat Greek Strained Yogurt With Blueberry}	=> {Total 2% with Strawberry Lowfat Greek Strained Yogurt}	0.002902447	0.4495134	0.006456863	48.34339	9331
[8]	{Total 2% Lowfat Greek Strained Yogurt with Peach}	=> {Total 2% with Strawberry Lowfat Greek Strained Yogurt}	0.002492788	0.4025517	0.006192467	43.29285	8014
[9]	{Zero Calorie Cola}	=> {Soda}	0.001234885	0.4638934	0.002662002	41.66855	3970
[10]	{Sparkling Lemon Water, Sparkling Water Grapefruit}	=> {Lime Sparkling Water}	0.001379214	0.4807025	0.002869164	33.20152	4434

The rules with lower lift indicate a weak association between the items.

3. Frequent itemsets for departments in orders dataset

The support threshold is kept 0.2 to handle the highly frequent department itemsets.
The itemset min size is kept 2 as an item set should involve at least 2 items.

Apriori

Parameter specification:

confidence	minval	smax	arem	aval	originalSupport	maxtime	support	minlen	maxlen	target	ext
NA	0.1	1	none	FALSE	TRUE	5	0.2	2	10	frequent itemsets	TRUE

Frequent department itemsets ordered based on support: -

	items	support	transIdenticalToItemsets	count
[1]	{dairy eggs,produce}	0.5504835	0.041147491	1769735
[2]	{produce,snacks}	0.3372617	0.011079750	1084254
[3]	{beverages,produce}	0.3335636	0.015122832	1072365
[4]	{dairy eggs,snacks}	0.3225479	0.006419225	1036951
[5]	{beverages,dairy eggs}	0.3207379	0.011243053	1031132
[6]	{frozen,produce}	0.2981184	0.008441077	958413
[7]	{pantry,produce}	0.2869403	0.008534394	922477
[8]	{dairy eggs,frozen}	0.2833797	0.004385553	911030
[9]	{dairy eggs,produce,snacks}	0.2704660	0.016170152	869514
[10]	{dairy eggs,pantry}	0.2692616	0.004630353	865642
[11]	{beverages,dairy eggs,produce}	0.2611630	0.021223227	839606
[12]	{dairy eggs,frozen,produce}	0.2425800	0.014641320	779864
[13]	{dairy eggs,pantry,produce}	0.2324586	0.014968860	747325
[14]	{beverages,snacks}	0.2291035	0.009489330	736539
[15]	{bakery,produce}	0.2289679	0.004378088	736103
[16]	{bakery,dairy eggs}	0.2249973	0.002809752	723338
[17]	{deli,produce}	0.2060591	0.003907463	662454

- With a support threshold of 0.2 there are 17 paired departments.
- People buy dairy eggs often. 10 out of the 17 pairs with highest support contain dairy eggs.
- All of the items are edibles.

4. Association rules for departments in orders dataset

A support threshold 0.1 is used for less frequent items with a high confidence 0.6 to generate strong rules.

set of 94 rules

```
rule length distribution (lhs + rhs):sizes
 1  2  3  4
 2 24 47 21
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.000	2.000	3.000	2.926	3.000	4.000

summary of quality measures:

support	confidence	coverage	lift	count
Min. :0.1014	Min. :0.6002	Min. :0.1144	Min. :0.9272	Min. : 326099
1st Qu.:0.1176	1st Qu.:0.8103	1st Qu.:0.1406	1st Qu.:1.1342	1st Qu.: 377939
Median :0.1407	Median :0.8392	Median :0.1749	Median :1.1768	Median : 452328
Mean :0.1824	Mean :0.8261	Mean :0.2263	Mean :1.1769	Mean : 586505
3rd Qu.:0.1946	3rd Qu.:0.8658	3rd Qu.:0.2370	3rd Qu.:1.2261	3rd Qu.: 625603
Max. :0.7494	Max. :0.9040	Max. :1.0000	Max. :1.4023	Max. :2409320

mining info:

data	ntransactions	support	confidence
transactions	3214874	0.1	0.6

Ordered by Confidence

	lhs	rhs	support	confidence	coverage	lift	count
[1]	{canned goods,dairy eggs}	=> {produce}	0.1536741	0.9040459	0.1699849	1.206313	494043
[2]	{dairy eggs,meat seafood}	=> {produce}	0.1291578	0.9014228	0.1432821	1.202813	415226
[3]	{dairy eggs,deli,snacks}	=> {produce}	0.1042573	0.8996680	0.1158842	1.200471	335174
[4]	{dairy eggs,dry goods pasta}	=> {produce}	0.1377208	0.8942269	0.1540110	1.193211	442755
[5]	{dairy eggs,frozen,pantry}	=> {produce}	0.1175595	0.8898587	0.1321103	1.187382	377939
[6]	{dairy eggs,deli}	=> {produce}	0.1717775	0.8897602	0.1930604	1.187251	552243
[7]	{bakery,frozen,produce}	=> {dairy eggs}	0.1014345	0.8868326	0.1143784	1.309422	326099
[8]	{deli,frozen}	=> {produce}	0.1034722	0.8867569	0.1166861	1.183243	332650
[9]	{bakery,produce,snacks}	=> {dairy eggs}	0.1126784	0.8840252	0.1274607	1.305277	362247
[10]	{bakery,dairy eggs,frozen}	=> {produce}	0.1014345	0.8837611	0.1147759	1.179246	326099

Ordered by Lift

	lhs	rhs	support	confidence	coverage	lift	count
[1]	{dairy eggs,deli,produce}	=> {snacks}	0.1042573	0.6069321	0.1717775	1.402289	335174
[2]	{beverages,frozen}	=> {snacks}	0.1123067	0.6021769	0.1865012	1.391302	361052
[3]	{dairy eggs,deli}	=> {snacks}	0.1158842	0.6002481	0.1930604	1.386846	372553
[4]	{bakery,frozen,produce}	=> {dairy eggs}	0.1014345	0.8868326	0.1143784	1.309422	326099
[5]	{bakery,produce,snacks}	=> {dairy eggs}	0.1126784	0.8840252	0.1274607	1.305277	362247
[6]	{bakery,beverages,produce}	=> {dairy eggs}	0.1034009	0.8773294	0.1178587	1.295391	332421
[7]	{bakery,pantry}	=> {dairy eggs}	0.1073034	0.8668384	0.1237871	1.279901	344967
[8]	{deli,produce,snacks}	=> {dairy eggs}	0.1042573	0.8661252	0.1203721	1.278848	335174
[9]	{bakery,frozen}	=> {dairy eggs}	0.1147759	0.8658668	0.1325561	1.278466	368990
[10]	{breakfast,produce}	=> {dairy eggs}	0.1140661	0.8651385	0.1318472	1.277391	366708

The rules with higher lift indicate a stronger association between the departments

A higher support threshold 0.2 is used for more frequent items with a lower confidence 0.4 to generate weaker rules.

set of 36 rules

rule length distribution (lhs + rhs):sizes

```
1 2 3
4 20 12
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.000	2.000	2.000	2.222	3.000	3.000

summary of quality measures:

support	confidence	coverage	lift	count
Min. :0.2061	Min. :0.4184	Min. :0.2396	Min. :0.9819	Min. : 662454
1st Qu.:0.2426	1st Qu.:0.4758	1st Qu.:0.3308	1st Qu.:1.0447	1st Qu.: 779864
Median :0.2769	Median :0.7473	Median :0.4431	Median :1.1076	Median : 890272
Mean :0.3223	Mean :0.6732	Mean :0.5135	Mean :1.1041	Mean :1036023
3rd Qu.:0.3336	3rd Qu.:0.8130	3rd Qu.:0.6773	3rd Qu.:1.1530	3rd Qu.:1072365
Max. :0.7494	Max. :0.8633	Max. :1.0000	Max. :1.2144	Max. :2409320

mining info:

data	ntransactions	support	confidence
transactions	3214874	0.2	0.4

Ordered by Confidence

	lhs	rhs	support	confidence	coverage	lift	count
[1]	{dairy eggs,pantry}	=> {produce}	0.2324586	0.8633188	0.2692616	1.151969	747325
[2]	{deli}	=> {produce}	0.2060591	0.8599948	0.2396050	1.147533	662454
[3]	{dairy eggs,frozen}	=> {produce}	0.2425800	0.8560245	0.2833797	1.142236	779864
[4]	{dairy eggs,snacks}	=> {produce}	0.2704660	0.8385295	0.3225479	1.118891	869514
[5]	{bakery}	=> {produce}	0.2289679	0.8350042	0.2742117	1.114187	736103
[6]	{pantry}	=> {produce}	0.2869403	0.8251933	0.3477250	1.101096	922477
[7]	{bakery}	=> {dairy eggs}	0.2249973	0.8205242	0.2742117	1.211517	723338
[8]	{beverages,dairy eggs}	=> {produce}	0.2611630	0.8142566	0.3207379	1.086503	839606
[9]	{frozen,produce}	=> {dairy eggs}	0.2425800	0.8137035	0.2981184	1.201446	779864
[10]	{dairy eggs}	=> {produce}	0.5504835	0.8127976	0.6772701	1.084556	1769735

Ordered by Lift

	lhs	rhs	support	confidence	coverage	lift	count
[1]	{dairy eggs,produce}	=> {pantry}	0.2324586	0.4222807	0.5504835	1.214410	747325
[2]	{bakery}	=> {dairy eggs}	0.2249973	0.8205242	0.2742117	1.211517	723338
[3]	{frozen,produce}	=> {dairy eggs}	0.2425800	0.8137035	0.2981184	1.201446	779864
[4]	{dairy eggs,produce}	=> {frozen}	0.2425800	0.4406671	0.5504835	1.199549	779864
[5]	{pantry,produce}	=> {dairy eggs}	0.2324586	0.8101286	0.2869403	1.196168	747325
[6]	{produce,snacks}	=> {dairy eggs}	0.2704660	0.8019468	0.3372617	1.184087	869514
[7]	{snacks}	=> {beverages}	0.2291035	0.5293331	0.4328154	1.167694	736539
[8]	{beverages}	=> {snacks}	0.2291035	0.5053957	0.4533151	1.167694	736539
[9]	{beverages,produce}	=> {dairy eggs}	0.2611630	0.7829480	0.3335636	1.156035	839606
[10]	{dairy eggs,pantry}	=> {produce}	0.2324586	0.8633188	0.2692616	1.151969	747325

The rules with lower lift indicate a weak association between the departments.