Market basket analysis

Agata Załęska, Mikołaj Szkaradek, Zofia Ogonek

Introduction

- Market Basket Analysis - one of the key techniques for uncovering associations between items
- The main goal - understanding customer behaviour
- The received data can be used to increase efficiency of sales and marketing

The data

- The data contains information about 3777580 transactions from 2011-01-01 to 2014-10-01
- There are 8159538 records in total
- There is information about 204435 clients (some of the transactions have unknown clients)
- Some of the transactions have negative price (probably returns) - we decided not to take them into account

	transaction_id	sales_datetime	customer_id
0	2e18343f9b9a95e89587273536e59d6e	2011-01-01 09:04:00	-1
1	d53096e90b515b563631b18acfa4d364	2011-01-01 09:04:00	-1
2	d53096e90b515b563631b18acfa4d364	2011-01-01 09:04:00	-1

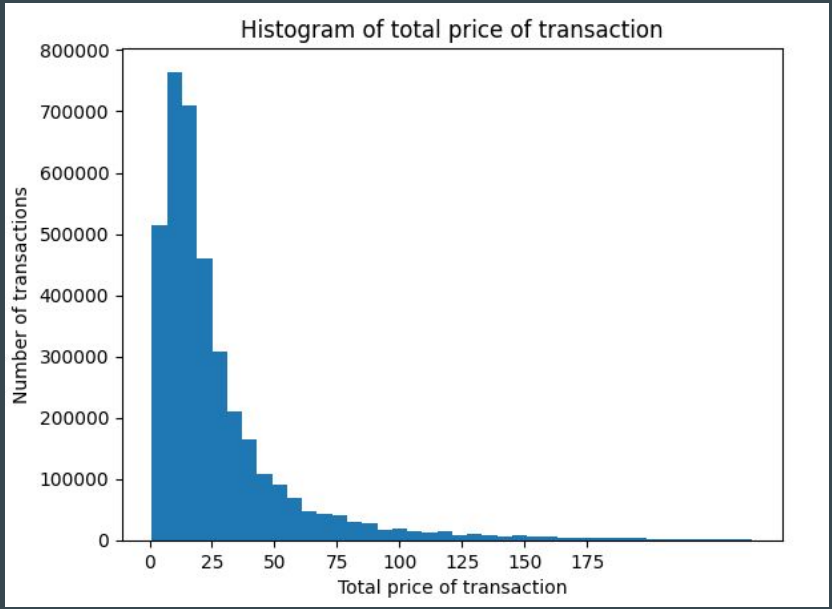
	product_id	quantity	price	category_id	parent_id
0	006ae35b6f0aae363ff038ffd44ad049	1.0	13.5	208	29
1	780024e0152928b310df607663294dd4	1.0	6.5	179	30
2	22981f293030ae132845164a0ba728e4	1.0	6.5	179	30

	store_id	department_id	salesperson_id
0	18	2	108
1	17	2	108
2	17	2	108

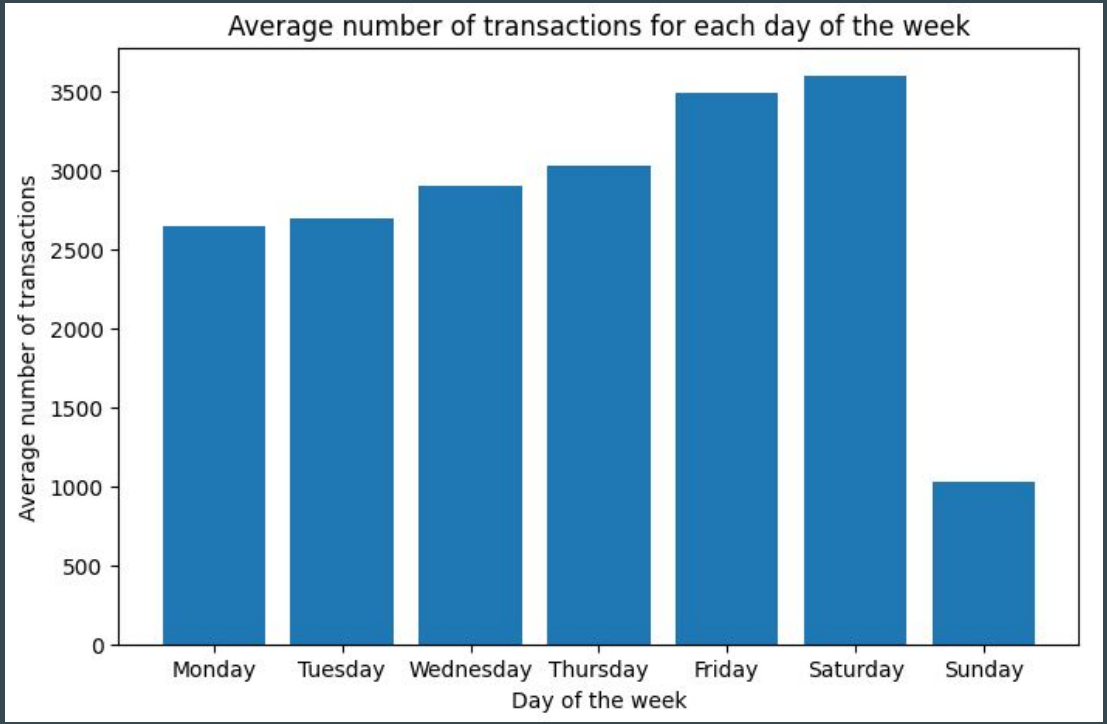
```
transaction_id
sales_datetime
customer_id
product_id
quantity
price
category_id
parent_id
store_id
department_id
salesperson_id
```

Transactions

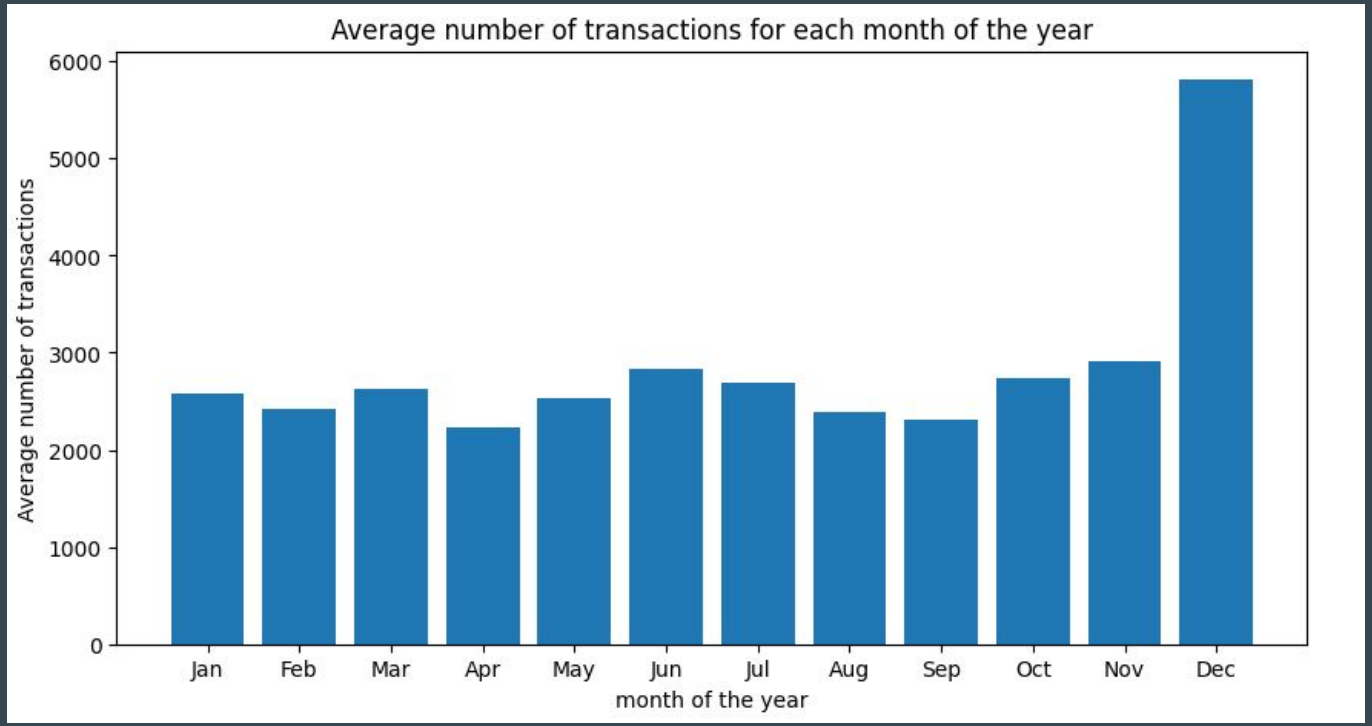
- Mean total price of transaction is 29.544240753418883
- Median total price of transaction is 18.0
- Standard deviation of total price of transaction is 42.31757797922174
- On average there are 2774 transactions per day



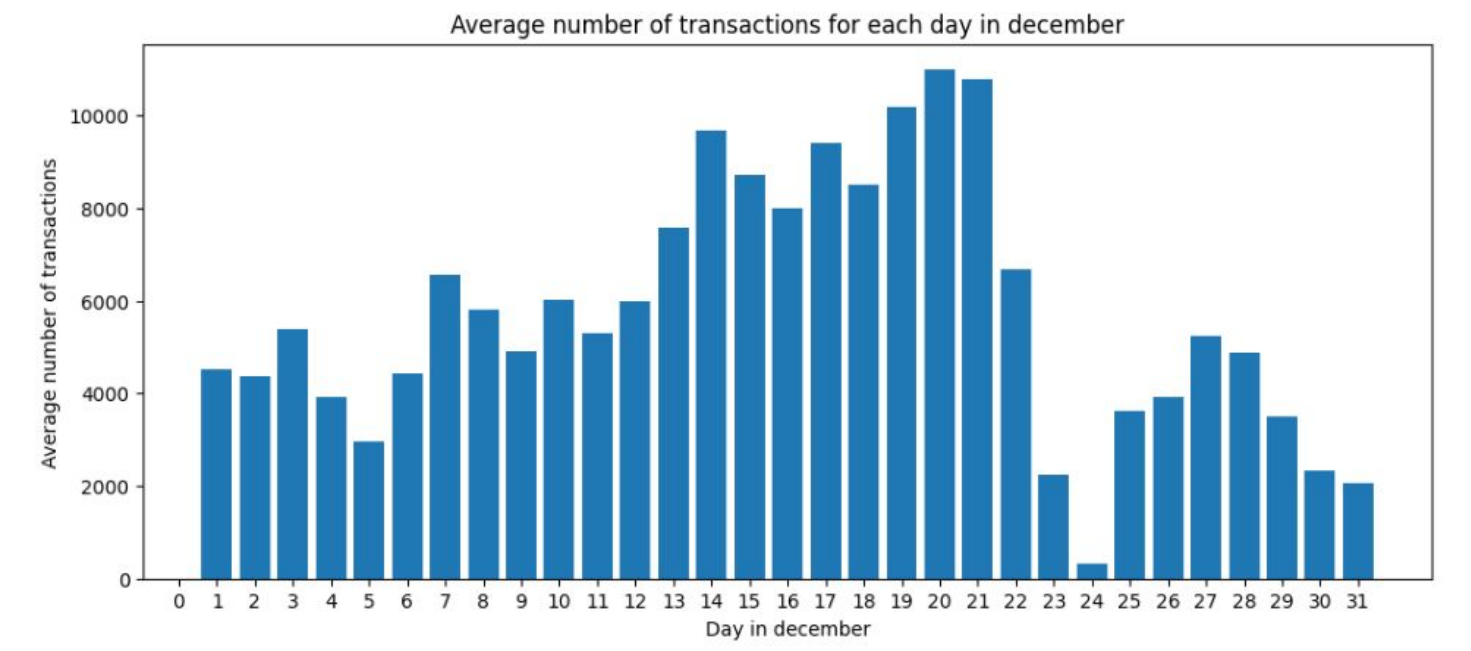
When customers buy the most?



When customers buy the most?

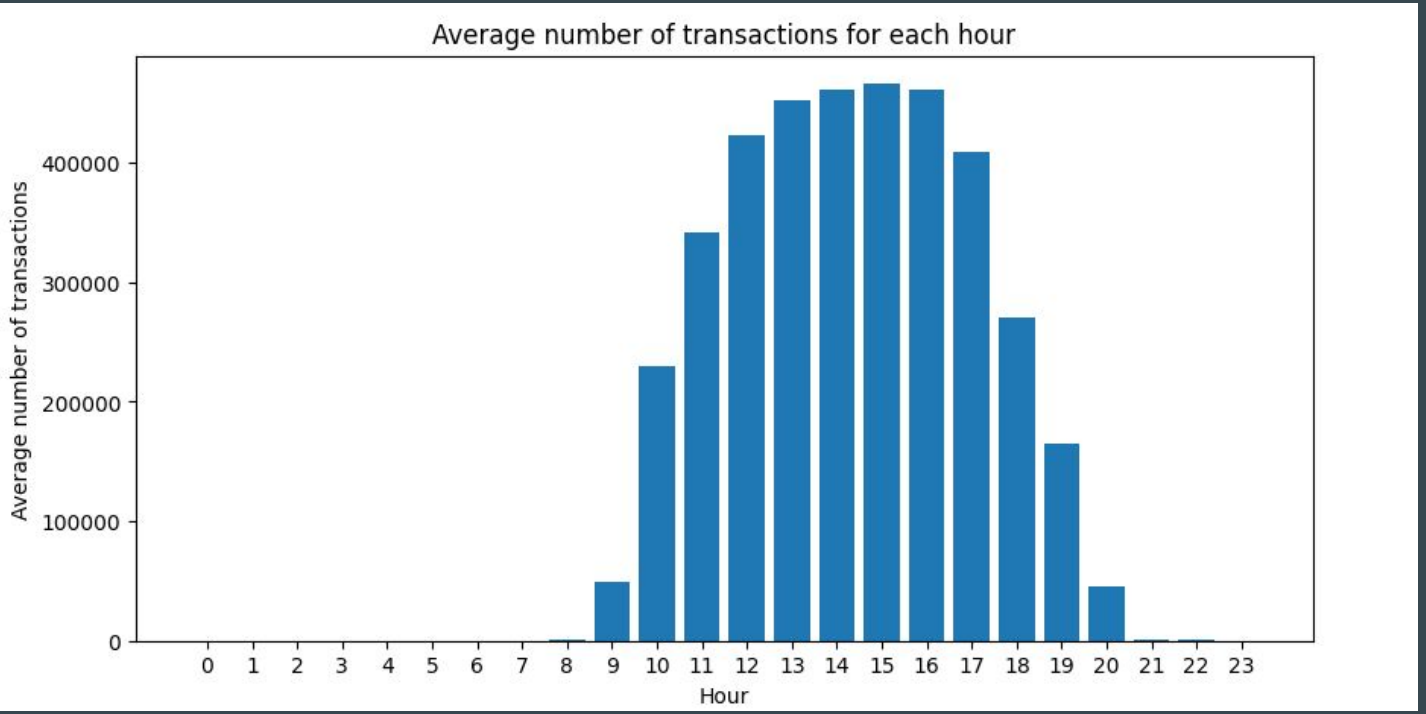


December


Average number of transactions for each day in december
Average number of transactions
0
2000
4000
6000
8000
10000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
Day in december

What time do customers buy the most?

	hour	number_of_transactions
0	0	13
1	1	6
2	2	25
3	3	42
4	4	81
5	5	101
6	6	118
7	7	202
8	8	1066
9	9	49724
10	10	230101
11	11	341152
12	12	423028
13	13	451728
14	14	461010
15	15	466166
16	16	461261
17	17	408910
18	18	270182
19	19	165146
20	20	45576
21	21	1414
22	22	372
23	23	156



Association rules

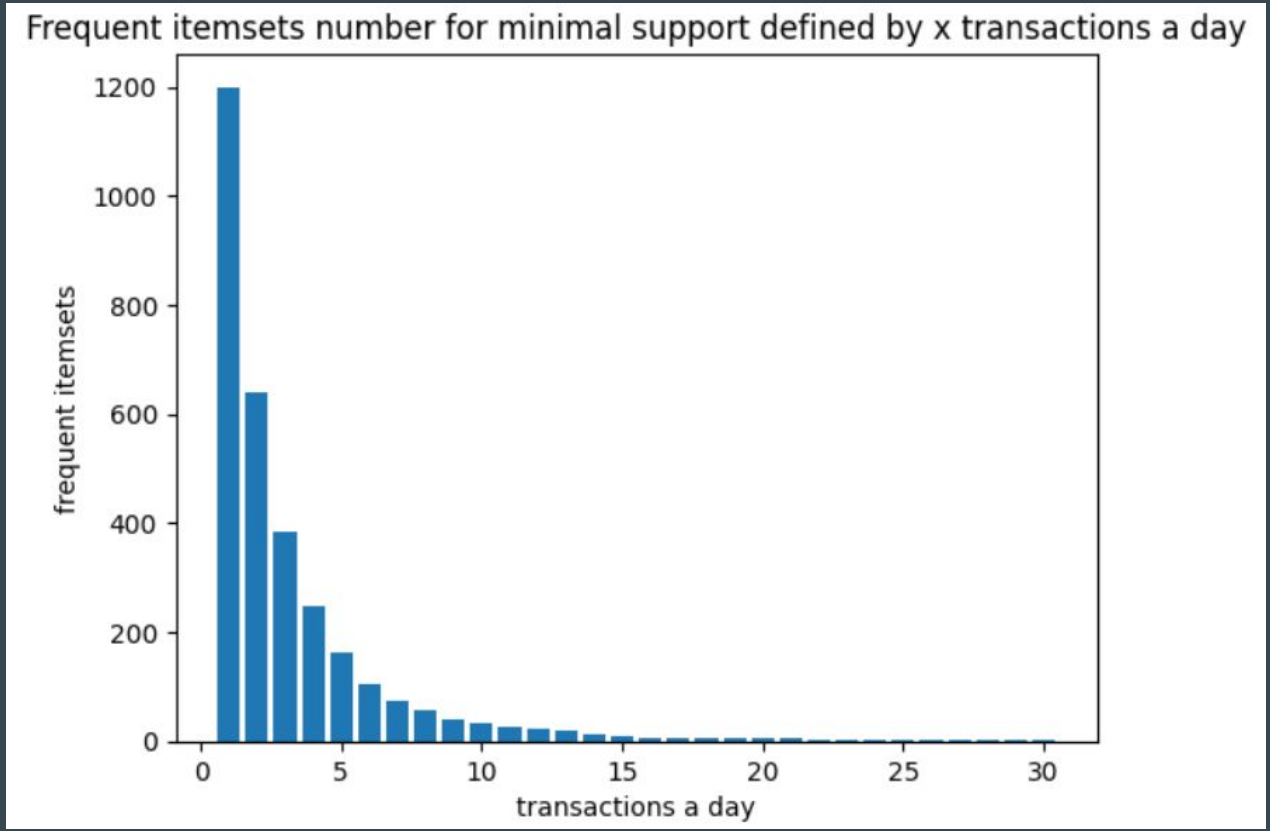
A sparse matrix

	0	1	2	3	4	5	6	7	8	9	...	\
0	False	False	False	False	False	False	False	False	False	False	...	
1	False	False	False	False	False	False	False	False	False	False	...	
2	False	False	False	False	False	False	False	False	False	False	...	
3	False	False	False	False	False	False	False	False	False	False	...	
4	False	False	False	False	False	False	False	False	False	False	...	

	94092	94093	94094	94095	94096	94097	94098	94099	94100	94101
0	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False

Minimal support

- The first step is to define minimal support (what do we mean by “frequent” itemset)
- We decided to calculate the minimal support using the formula `min_support = (x * days_count) / transactions_count`
- Example: for items bought approximately 10 times a day the support is about 0.003621
- We also had to ensure that there is going to be a reasonable number of frequent itemsets for a given minimal support


Frequent itemsets number for minimal support defined by x transactions a day
frequent itemsets
1200
1000
800
600
400
200
0
0
5
10
15
20
25
30
transactions a day

Apriori vs FP-growth

- Unfortunately, the results were unsatisfactory - for minimal support smaller than (x * days_count) / transactions_count we received “dead kernel” error
- For the smallest value of minimal support without error, we’ve only got about 40 frequent itemsets
- FP-growth algorithm worked much better

FP-growth algorithm - association rules

- We calculated association rules for lift > 1 and minimal support = 2*days_count/transactions_count , for different values of minimal confidence
- Example results:

there are 30 association rules

	antecedents	consequents	antecedent support	consequent support	support
0	(61987)	(28829)	0.004929	0.013583	0.000945
1	(28829)	(61987)	0.013583	0.004929	0.000945
2	(6427)	(28829)	0.003674	0.013583	0.000972
3	(28829)	(6427)	0.013583	0.003674	0.000972
4	(28829)	(54886)	0.013583	0.003615	0.000789
5	(54886)	(28829)	0.003615	0.013583	0.000789
6	(73032)	(28829)	0.008057	0.013583	0.001337
7	(28829)	(73032)	0.013583	0.008057	0.001337
8	(73032)	(87313)	0.008057	0.011456	0.000783
9	(87313)	(73032)	0.011456	0.008057	0.000783
10	(31216)	(28829)	0.004557	0.013583	0.000774
11	(28829)	(31216)	0.013583	0.004557	0.000774
12	(44584)	(28829)	0.004552	0.013583	0.001008
13	(28829)	(44584)	0.013583	0.004552	0.001008
14	(67387)	(28829)	0.004247	0.013583	0.000916
15	(28829)	(67387)	0.013583	0.004247	0.000916

Association rules - grouping by store_id

- We've noticed that the rules we received had pretty low support
- To try improving that, we tried filtering the data by store (maybe different stores sell different products, hence the low support)
- The results were quite weird and unexpected
- For the same minimal support as for the whole data, most of the stores got a similar number of rules
- The results for different shops were drastically different

```
For store with id = 38:
There are 5249 frequent itemsets
There are 179546 association rules
Max support is 0.06243567753001715
```

```
For store with id = 1:
There are 403 frequent itemsets
There are 12 association rules
Max support is 0.0016897547501464802
```

```
For store with id = 39:
There are 734 frequent itemsets
There are 54 association rules
Max support is 0.0016319869441044472
```

```
For store with id = 2:
There are 660 frequent itemsets
There are 40 association rules
Max support is 0.002082108563590045
```

```
For store with id = 40:
There are 699 frequent itemsets
There are 170 association rules
Max support is 0.004967288587351587
```

min_support = (x * days_count) / transactions_in_given_store_count.

Grouping by category_id approach

- We excluded categories that have less than 15 products .
- We also excluded categories that appear in less than 20 transactions a day.
- There were 57 such categories.

- The target was to find frequent itemsets among 1 category or pair of 2 categories.
- We grouped data so that the operations on it were cheaper and faster.

Results of category split approach

To find associations rules among one category we created sparse matrix with transactions containing only products from this category and used FP-growth algorithm

```
Analysing category: 179
Number of transactions in this category is 135255
Number of different products in this category is 140
min support is 0.010069868027059998
there are 54 itemsets that appear in at least 1 transactions of this category a day
maximum support is 0.0347935381316772
there are 14 association rules with lift at least 1 and support > 0.01
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(59)	(52)	0.024731	0.034298	0.013582	0.549178	16.011868	0.012734	2.142091
1	(52)	(59)	0.034298	0.024731	0.013582	0.395991	16.011868	0.012734	1.614658
2	(25)	(66)	0.077188	0.058815	0.034794	0.450766	7.664160	0.030254	1.713633
3	(66)	(25)	0.058815	0.077188	0.034794	0.591578	7.664160	0.030254	2.259456
4	(43)	(14)	0.033936	0.022084	0.010373	0.305664	13.840861	0.009624	1.408420
5	(14)	(43)	0.022084	0.033936	0.010373	0.469702	13.840861	0.009624	1.821738
6	(16)	(37)	0.025751	0.035001	0.017752	0.689348	19.695353	0.016850	3.106371
7	(37)	(16)	0.035001	0.025751	0.017752	0.507182	19.695353	0.016850	1.976894
8	(81)	(31)	0.031311	0.028420	0.015497	0.494923	17.414372	0.014607	1.923628
9	(31)	(81)	0.028420	0.031311	0.015497	0.545265	17.414372	0.014607	2.130229
10	(121)	(19)	0.026565	0.042564	0.016406	0.617590	14.509658	0.015275	2.503688

Category pair result

We've achieved similar results using category pairs, so this did not help us with finding something new and interesting.

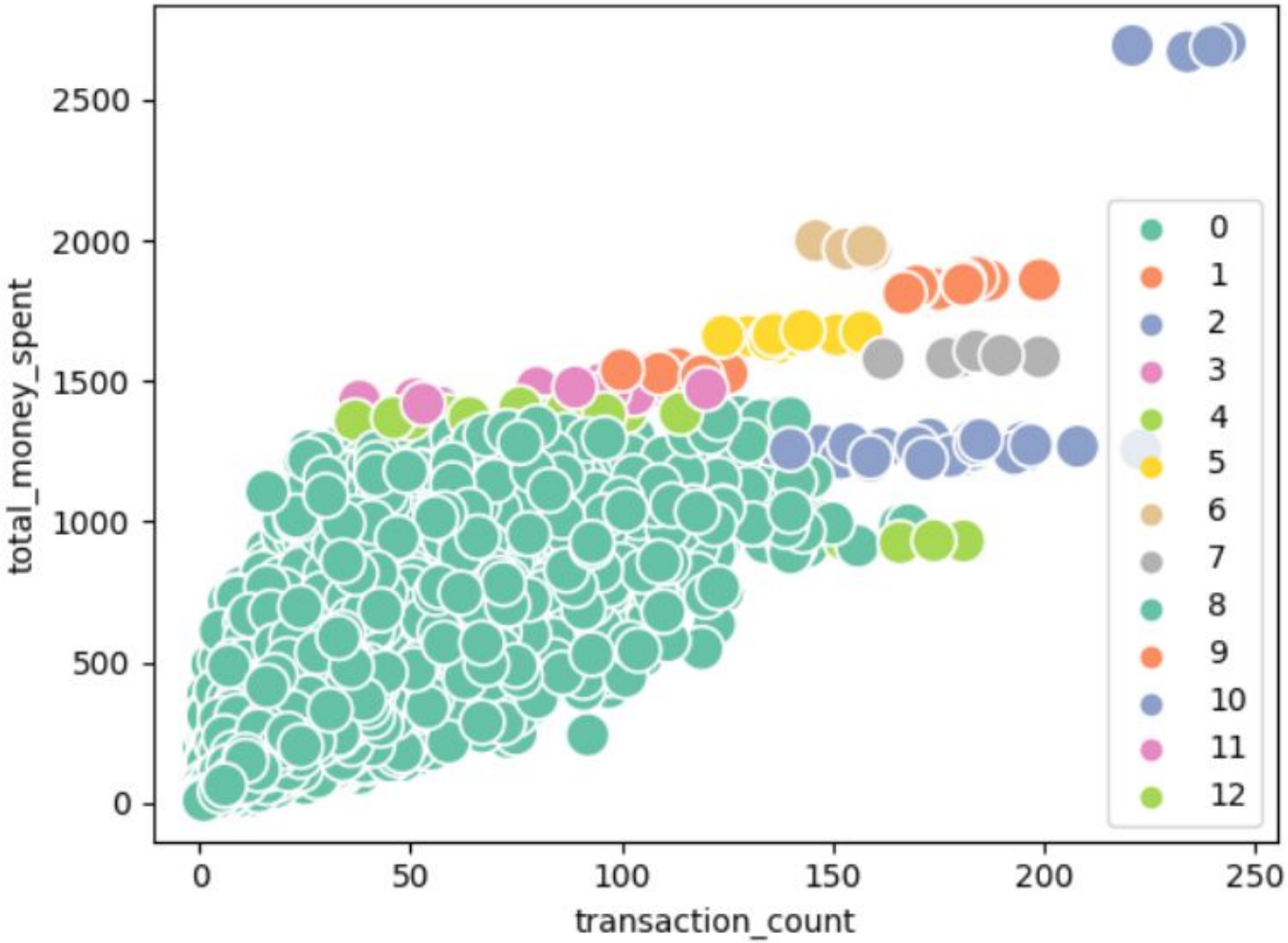
```
Analysing category pair: 208 179
Number of transactions in this pair is 217790
Number of different products in this pair is 164
min support is 0.006253730657973277
there are 71 itemsets that appear in at least 1 transactions of this category a day
maximum support is 0.021607970981220442
there are 6 association rules with lift at least 1 and support > 0.01
   antecedents consequents  antecedent support  consequent support   support  \
4         (27)        (78)            0.047936            0.036526  0.021608
5         (78)        (27)            0.036526            0.047936  0.021608
10        (18)        (44)            0.015992            0.021737  0.011024
11        (44)        (18)            0.021737            0.015992  0.011024
14        (21)       (142)            0.026434            0.016498  0.010189
15       (142)        (21)            0.016498            0.026434  0.010189
```

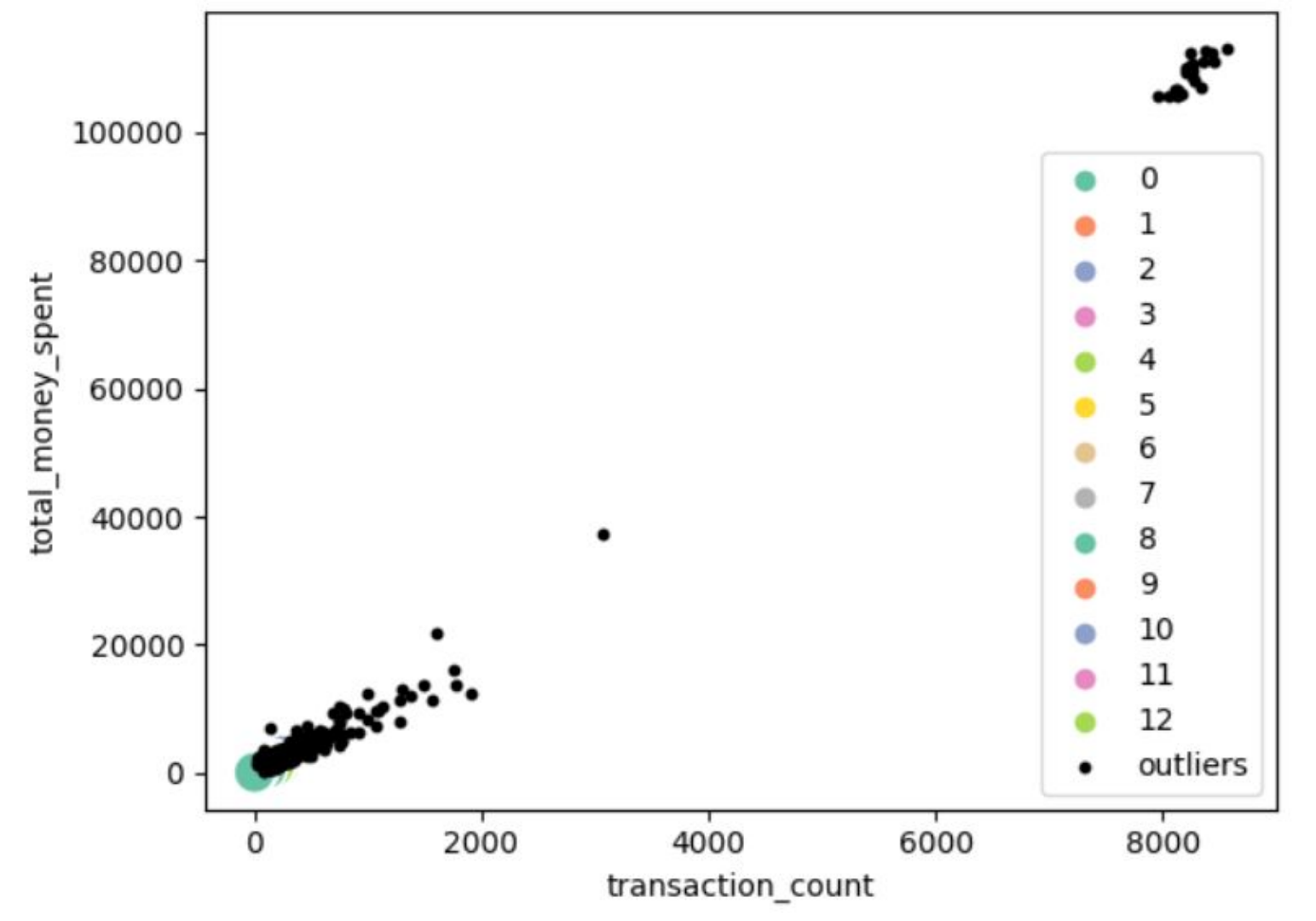
Customer analysis

DB-scan algorithm

- We used DB-Scan algorithm to cluster the customers into similar groups
- How to define “similar”?
- The first attempt:
  - For each customer we calculated the number of their transactions and the total amount of money they spent
  - Based on this information, we divided them into groups of those who have similar purchasing habits
- We used the data about 35 000 customers, as the algorithm could not handle the data about all of them (about 200 000)

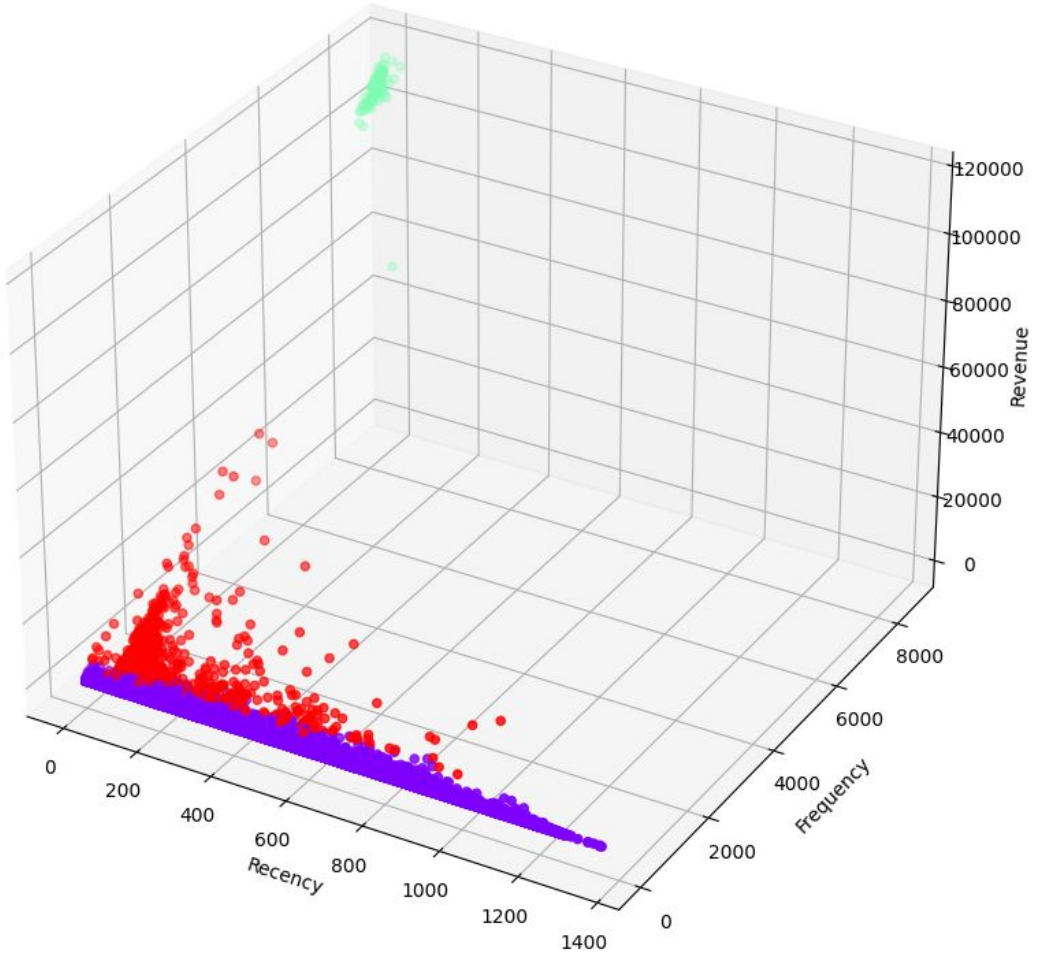
customer_id	transaction_count	total_money_spent
0	9	252.99
1	6	20.50
2	40	410.80
3	1	10.00
4	6	60.87


2500
2000
1500
1000
500
0
total_money_spent
0
50
100
150
200
250
transaction_count
0
1
2
3
4
5
6
7
8
9
10
11
12


total_money_spent
100000
80000
60000
40000
20000
0
0
2000
4000
6000
8000
transaction_count
0
1
2
3
4
5
6
7
8
9
10
11
12
outliers

Customer lifecycle analysis

- There are 110420 acquisition customers - customers who meet the criteria for being new customers, typically having made few purchases and having recently made their first purchase
- There are 1368 retention customers - customers who meet the criteria for being loyal customers, typically having made multiple purchases and having recently made a purchase
- There are 92653 churn customers - customers who meet the criteria for being at risk of leaving the business, typically having made few purchases or not having made a purchase in a while


120000
100000
80000
60000
40000
20000
0
Revenue
8000
6000
4000
2000
0
Frequency
0
200
400
600
800
1000
1200
1400
Recency

Thank you for your attention

...