Modelling supermarket transaction data

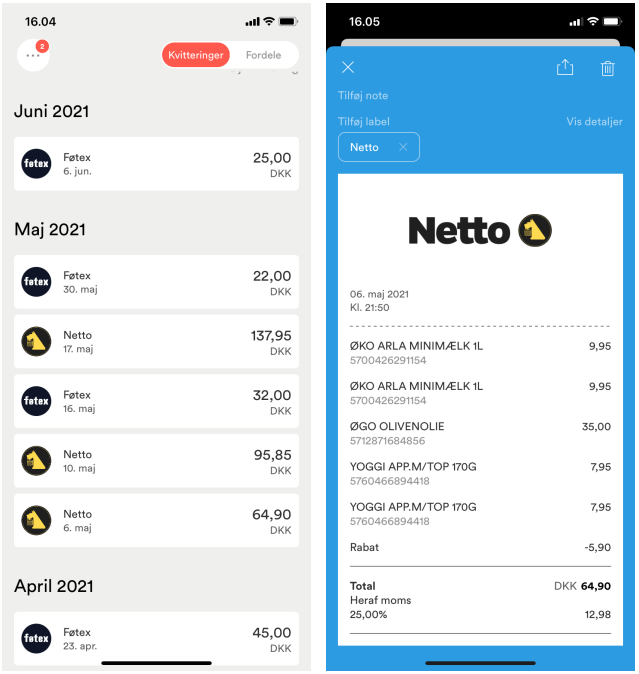
Emilie Prang Nielsen

June 09, 2021

About me

- Master in mathematics-economics in 2018.
- 1,5 years in Novo Nordisk as biostatistics graduate.
- Travelled and studied in 2020 due to "Rejselegat for Matematikere".
- Started industrial PhD in the spring 2021 at Hjerteforeningen with Thomas Gerds as university supervisor.
- Couple of days a week each place, same office as Anne and Aksel when at biostat :)

Working hypothesis: Show me your grocery shopping basket and I will tell you who you are



- The storebox app.
- SMIL study: register on storebox with CPR-number -> connect registries (age, sex, prescriptions etc.) with electronic receipts, encrypted.
- Around 9,000 signed up: time series of grocery shopping from around 2 years (fall 2018 to fall 2020).
- Project possible due to Christian Torp. Kathrine Kold (master in public health) also on storebox data.

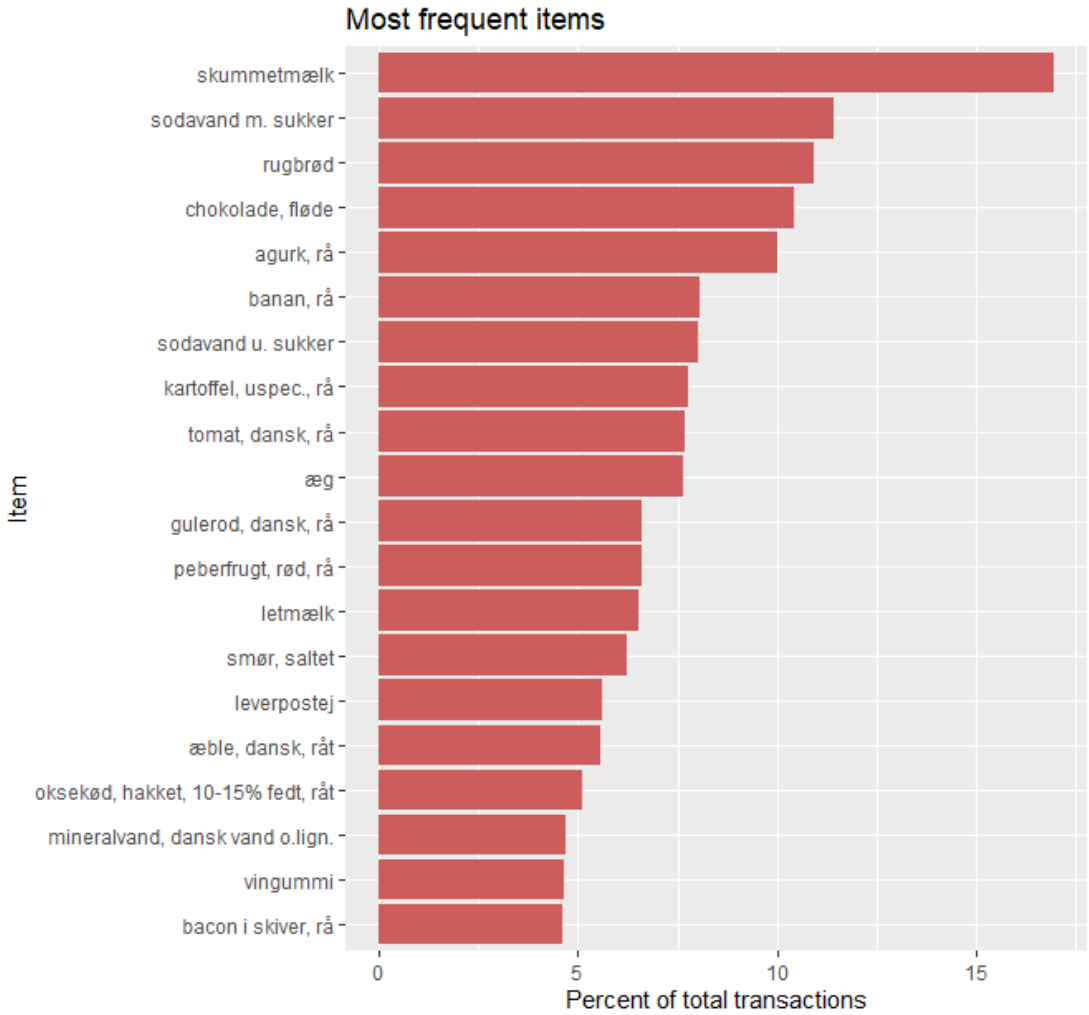
Supermarket transaction data structure

- Each transaction is uniquely defined by person and time, and contains at least one item.
- Almost 1 million transactions in total.
- One can think of a transaction as a receipt.

TID	person	time	item	itemprice	quantity	transactionprice
1	1	17-03-2019 08:03:00	bread	11.95	1	76.95
1	1	17-03-2019 08:03:00	dip	6.00	2	76.95
1	1	17-03-2019 08:03:00	dressings	53.00	1	76.95
2	1	19-03-2019 10:15:53	fresh eggs	27.95	1	78.40
2	1	19-03-2019 10:15:53	apples	15.00	0.700	78.40
2	1	19-03-2019 10:15:53	dip	10.00	2	78.40
2	1	19-03-2019 10:15:53	bread	19.95	1	78.40
3	2	02-02-2020 19:34:01	milk	9.95	1	9.95
4	2	14-02-2020 15:55:04	wine	109.00	3	479.80
4	2	14-02-2020 15:55:04	beef	49.95	2	479.80
4	2	14-02-2020 15:55:04	yoghurt	18.95	1	479.80
4	2	14-02-2020 15:55:04	bread	5.00	5	479.80
4	2	14-02-2020 15:55:04	milk	8.95	1	479.80
5	2	20-02-2020 20:24:10	apples	2.00	2	19.00
5	2	20-02-2020 20:24:10	bread	15.00	1	19.00
...	...	...	...	...	...	

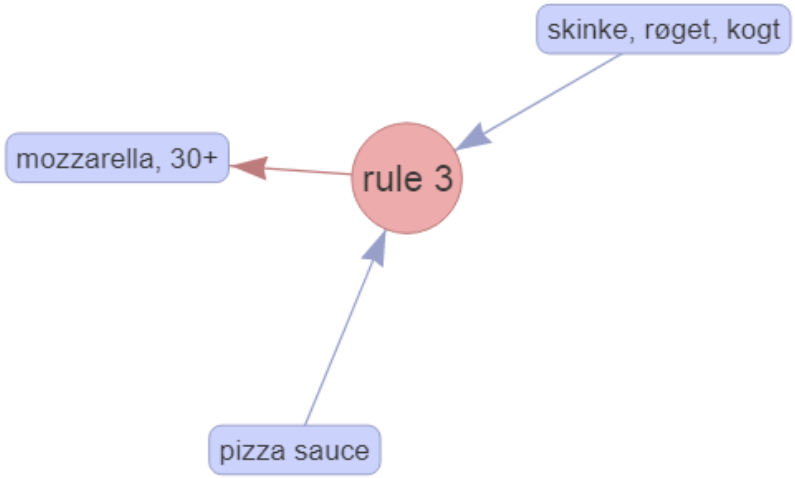
Categorization of purchases

- DTU: frida food database.
- Around 800 different food items (sorry for the Danish names!).



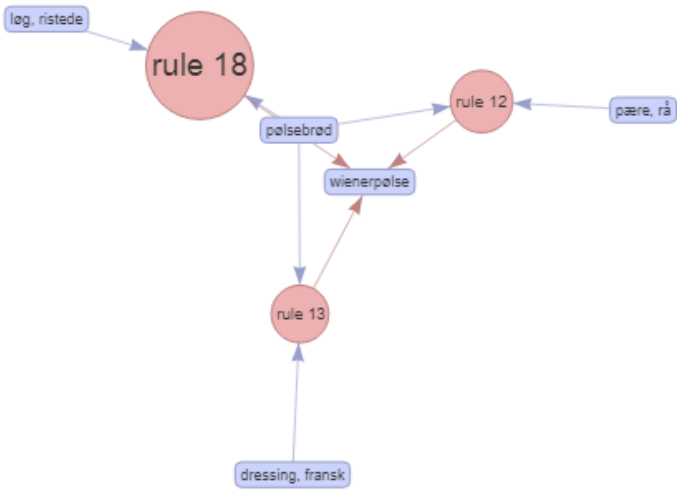
Exploratory: finding interesting patterns

- Association analysis: which items are bought frequently together, form association rules.
- A way for supermarkets to place their different items optimally to increase sales.
- For now: result of algorithm on storebox data with default settings (NOTE: don't ask about the details of the algorithm, used arules package in R, https://borgelt.net/apriori.html).
- Identify strong rules based on standard measures.

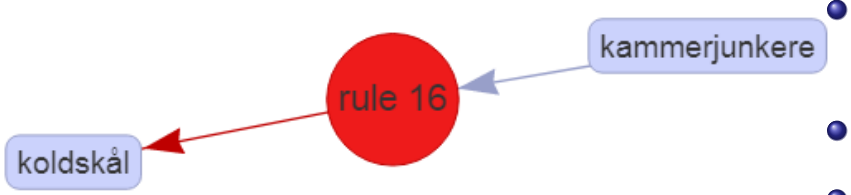


- Rule 3: {pizza sauce, ham} $\rightarrow$ {mozzarella}.
- Count: 209.
- Confidence: 80%.
- 209 transactions for the set and 80% of the times when you buy pizza sauce and ham you also buy mozzarella.

Association analysis continued



- Rule 18:
  {roasted onions, hotdog bread} → {sausage}.
- Count: 616.
- Confidence: 70%.
- Cluster of association rules that all connect to hotdogs.



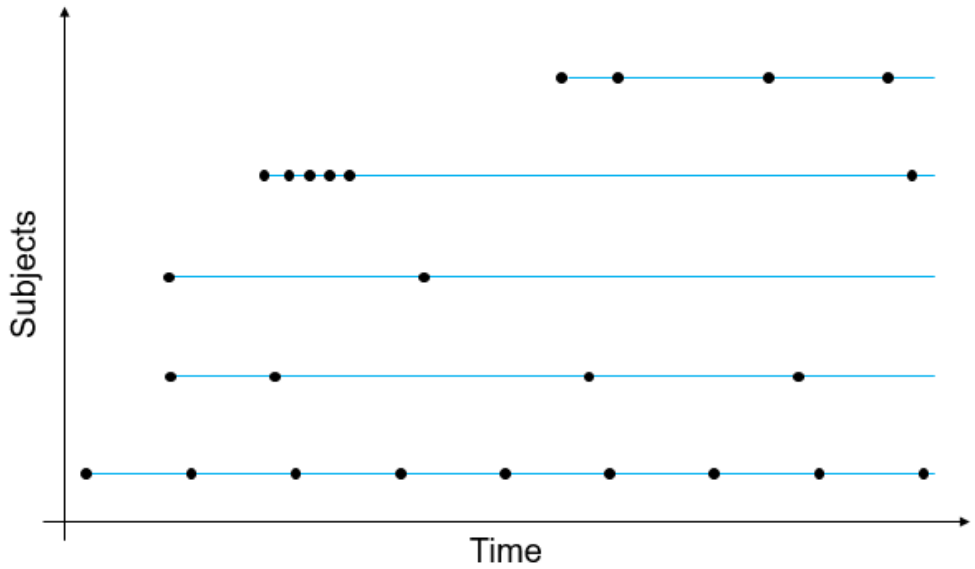
- Rule 16:
  {kammerjunkere (butter cookies)} → {koldskål}.
- Count: 3,660.
- Confidence: 40%.

- **Perspective:** Look more into the algorithm and expand this unsupervised approach to include labels (sex, age..) and time structure.

Possible research questions

- Does lockdown introduce a health risk? (for example for single citizens).
- Do younger people buy more heart healthy than older?
- Does a diabetes household shop differently than a non-diabetes household? (Kathrine).
- Need to incoorporate time structure and structural shocks (lockdown) to answer these.

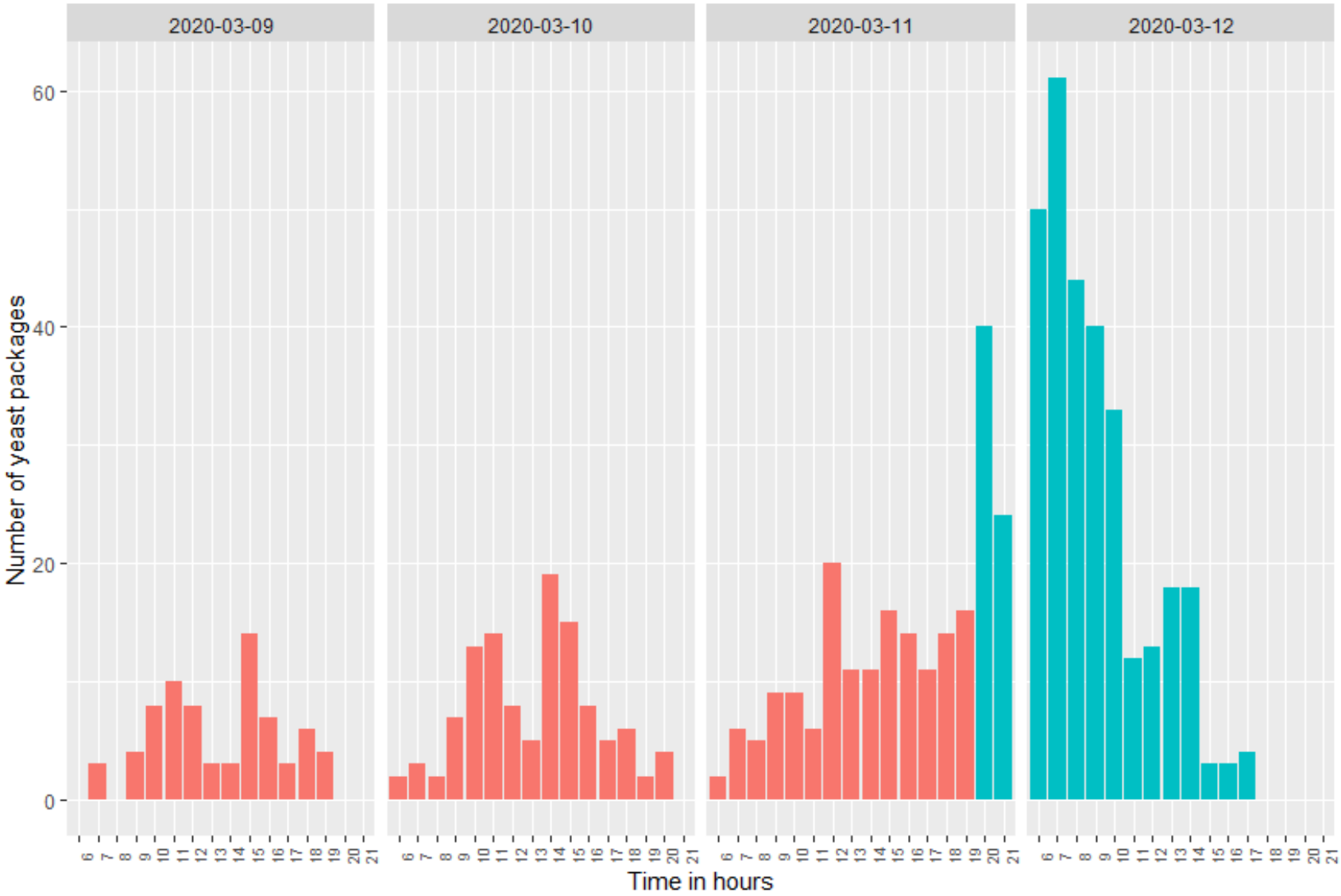
Complex time structure



- Delayed entry.
- Different frequency of transactions.
- Missing data in certain months.
- Missing information regarding shopping outside storebox realm (other supermarkets, greengrocers etc.)

Structural shock: lockdown

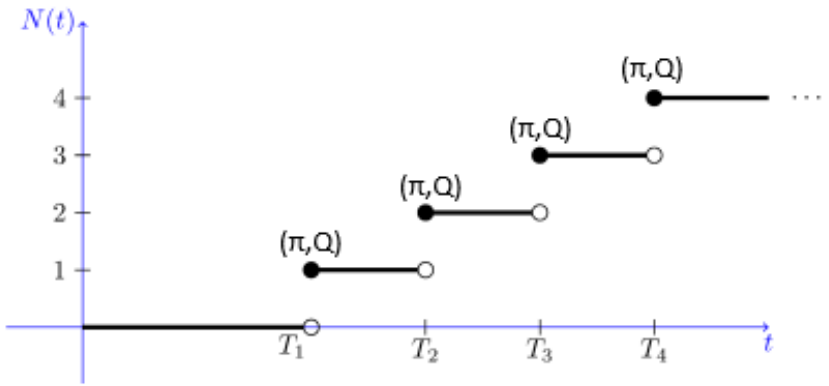
- Can we identify the supermarket run on yeast sales after the press meeting 11th March?



Mathematical framework

- Counting process, jumps at transaction times T_k. $N(t)$ is number of transactions in $[0, t]$.

$$N(t) = \sum_{k \geq 1} I\{T_k \leq t\}, \tag{1}$$



- Each item has corresponding price and quantity.
- Use framework for marked point processes, where we have a mark space for each item.
- For example for the item bread in the first transaction:
  $(\pi_{\text{bread}}(T_1), Q_{\text{bread}}(T_1)) = (11.95, 1) \in \mathcal{X}_{\text{bread}} = \mathbb{R}^+ \times \mathbb{R}^+$.

Today: Identify target parameters

- Current work: identify the target parameters for statistical analysis.
- Example 1: Expected number of transactions in [a,b] containing item m (for example monthly).

$$\begin{aligned}\mu^m(a,b) &= E\Big(\int_a^b f(X(t))dN(t)\Big) \\ &= \sum_{k=1}^{K} E(f(X(t))I_{\{a\leq T_k\leq b\}}) \\ &= \sum_{k=1}^{K} P(a\leq T_k\leq b, Q^m(T_k)>0, \pi^m(T_k)>0)\end{aligned}$$

where $f(X(t)) = I_{\{Q^m(t)>0,\ \pi^m(t)>0\}}$. So, $f(X(t))$ indicates the transactions where item m was purchased.

- Example 2: Expected amount spent on item m in [a,b]. Here, $f(X(t)) = Q^m(t)\cdot\pi^m(t)$.

Future work

- Need to adapt to missing data, different length of follow-up times and seasonality (important: covid).
- Modelling grocery transactions probabilities based on shopping history and patient characteristics (time series models, machine learning).
- Statistical inference for target parameters.