Vintereksamen - Dataanalyse

Januar 2023

Af Kenneth Gottfredsen, Eva Rauff og Sanne Sørensen

12/30/22

1 Problemfelt

Det har været et godt koldskålsår hos Thise Mejeri (Kjer 2022). Salget af koldskål er nemlig steget med 5%, sammenlignet med det forrige år (ibid). Sommervejret påvirker i stor grad kundernes efterspørgsel på koldskål (Kjer 2022, Jensen 2022, Holland 2022). Det er vanskeligt at forudsige præcist, hvor meget koldskål Thise skal producere til deres kunder. Hos Thise bruger medarbejderne vejrudsigten og mange års erfaring, når de skal vurdere, hvor mange liter koldskål, der skal produceres fremadrettet (Jensen 2022).

I en artikel fra TV Midtvest fortæller adm. direktør Poul Pedersen fra Thise Mejeri at: *"Vi kigger på tal og vejrudsigter som aldrig før. Så beslutter vi os for et tal, og vi plejer at være gode til at gætte." - (Jensen 2022).* Det er et erhvervsøkonomisk problem, hvis man udelukkende bruger vejrudsigten og gætteri til at forudsige, hvor meget koldskål produktionsafdelingen skal producere. Gætteriet indebærer en betydelig risiko, da man ikke kan garantere, at alt koldskålen bliver solgt - selv når vejret er godt! (Jensen 2022).

Konsekvensen kan føre til et stort økonomisk tab, fordi den mængde koldskål som ikke bliver solgt i butikkerne, leveres tilbage til Thises eget lager igen (Holland 2022) Afhænger COOPs efterspørgsel på koldskål kun af vejret? Eller er der andre mekanismer end vejret, som forårsager en stigende eller faldende efterspørgsel på koldskål?

For at løse dette problem bliver undersøgelsens formål derfor at bidrage med en multibel lineær regressionsmodel, som kan give en tilnærmelses forudsigelse af,

hvor mange liter koldskål COOP vil efterspørge. Den faglige vurdering er, at ovenstående problemstillinger sandsynligvis kan hænge sammen med Thises datamodenshedsproces. Det vil sige deres evne at bruge deres egne data til og skabe større økonomisk vækst i forbindelse med deres produktionsplanlægning.

Produktionsafdelingen hos Thise Mejeri mangler klarhed over, hvordan deres deres egen datamodenhedsproces hænger sammen med COOPs efterspørgsel af koldskål, samt hvilke andre mekanismer der kan have en positiv eller negativ indflydelse. Der findes forskellige teorier på området, som kan forklare sammenhængen. Det er disse teorier og andre vejr-variable, som er emnet for denne undersøgelse, da vi vil forsøge at forstå, hvordan disse hænger sammen med COOPs efterspørgsmål på koldskål, Thises datamodenshedsproces og andre faktorer.

1.1 Problemformulering

På baggrund af ovenstående problemfelt bliver der i næste afsnit formuleret et hovedspørgsmål og nogle underspørgsmål, de skal bruges til at besvare den erhvervsøkonomiske problemstillingen.

1.1.1 Hovedspørgsmål

"Hvordan kan Thises produktionsafdeling forbedre deres datamodenhed og dermed forbedre udnyttelsen af deres egne data til produktionsplanlægning?"

Hovedspørgsmålet er løsningsorienteret. Fordi spørgsmålet skal bidrage med datainitiativer som Thise selv kan benytte i sig af under produktionsplanlægningen.

Fordi datainitiativerne vil forbedre deres evne til at bruge data, den selv producerer, til og skabe større økonomisk værdi.

1.1.2 Underspørgsmål 1

"Hvor befinder Thises produktionsafdeling sig på nuværende tidspunkt datamodenhedsmæssigt?"

Underspørgsmålet skal skabe større forståelse for, hvor langt i datamodenhedsprocessen Thises produktionsafdeling befinder sig på nuværende tidspunkt.

1.1.3 Underspørgsmål 2

"Hvilke faktorer påvirker COOPs efterspørgsel af koldskål?"

Underspørgsmålet skal beskrive hvilke andre variable der hænger sammen med efterspørgslen af koldskål.

1.1.4 Underspørgsmål 3

"Hvilke faktorer kan Thises produktionsafdeling bruge til at forudsige COOPs efterspørgsel af koldskål?"

Underspørgsmålet skal analysere de bedste variable som produktionsafdeling skal have med i den endelige model, når de fremadrettet skal forsøge at forudsige Coops efterspørgsel af koldskål.

1.1.5 Underspørgsmål 4

"Kan undersøgelsens resultater anvendes til at styrke produktionsplanlægningen?" Underspørgsmålet skal perspektivere undersøgelsens statistiske analyse til en anden produktionskontekst.

2 Videnskabsteori

Et paradigme er et tankesystem (Bergfors 2021). Vores forståelse af sammenhængen mellem Thises datamodenhedsproces, vejrforholdene og efterspørgslen af koldskål er yderst kompleks. Der er behov for to forskellige former for viden om denne problemstilling. Vi har derfor valgt det realistiske paradigme, fordi vi gerne vil skabe større forståelse for, hvad der ligger bag Thises datamodenshedsproces. Vi vil også gerne kunne forklare med tal og dermed beskrive, hvorfor tallene ser ud, som de gør. Dette er, fordi vi antager, at Thises datamodenshedsproces hænger sammen med Coops efterspørgsmål på koldskål samt andre vejr-variables effekter herpå.

Med en *realistisk* vinkel kan man derfor anvende både kvalitative og kvantitative dataformer. Hvor datapunkter om efterspørgslen af koldskål fx består af tal og datapunkter om datamodenhedsprocessen fx omhandler det menneskelige sprog, holdninger, opfattelser. Vurderingen er, at de to former for data med fordel kan supplere hinanden i projektets analysedel. *Ontologi* er en antagelse om, hvordan man anskuer den verden, problemstillingen indgår i (Bergfors 2021).

Den ontologiske opfattelse er, at virkeligheden hos Thise Mejeri eksisterer udenfor eller inden i medarbejdere som arbejder på mejeriet. Men denne virkelighed opfattes som værende uafhængig af, hvordan en medarbejder opfatter verden, hvor sand viden om koldskål og datamodenhed i større grad er objektivt. (Bergfors 2021). Dette paradigme har derfor et større fokus på helheden i form af statistiske lovmæssigheder og repræsentativitet, men konteksten og enkeltdele spiller også en rolle. Havde vi kun fokus på tal, havde vi udelukkende valgt det positivistiske paradigme. Havde vi kun haft fokus på det enkelte menneskes sprog og opfattelser, havde vi valgt et socialkonstruktivistisk paradigme. Hvor verden udelukkende er en subjektiv social konstruktion, og hvor alt er under konstant forandring (ibid).

Epistemologi handler om, hvordan man anskaffer viden om en problemstilling (Bergfors 2021). Som nævnt tidligere vil vi benytte os af kvalitative og kvantitative data i samspil med hinanden, men at statistisk repræsentativitet spiller en større rolle i den her undersøgelse. Vores fokus bliver derfor på at forstå og fortolke, hvorfor tallene fremstår, som de gør, samt hvordan efterspørgslen af koldskål hænger sammen med Thises datamodenshedsproces set ud fra konteksten af produktionsafdelingen.

2.1 Undersøgelsesdesign

Et undersøgelsesdesign er en strategisk plan som skal besvare en problemstilling ud fra empiriske data (Bergfors 2021). Vi anvender et *statisk* undersøgelsesdesign til, at besvare vores problemstilling. Fordi designet i sig selv giver et her - og nu billede. Variablerne måles og observeres fx. fra perioden 1/4/22-30/8/22, hvil-

ket er et specifikt tidspunkt. Dertil undersøges relationen mellem vejrvariable og efterspørgslen på koldskål i den nuværende tilstand (ibid).

Designtypen er et *casestudium*, fordi vores problemstilling tager afsæt i en nutidig kontekst og fordi der skal produceres viden om komplekse sammenhænge (ibid). Casen omhandler Thise Mejeri som virksomhed. Vi vil derfor gerne vil beskrive og forstå det komplekse samspil som eksisterer mellem Thises datamodenhedsproces, vejret og COOPs efterspørgsel på koldskål.

2.2 Metodologi

Metodologi henviser til de forskellige metoder, man kan bruge til at indsamle data med (ibid). Vi har valgt at anvende eksisterende kvantitative data. Formålet med den kvantitative data er, at den skal bruges i forbindelse med en regressionsanalyse. Fordi vi gerne vil beskrive forskellige typer af vejr-variable - samt andre relevante variable påvirker Coops efterspørgsel af koldskål. Vi bruger eksisterende sekundære vejrdata hentet fra Danmarks meteorologiske instituttet (DMI), og produktionsdata om koldskål hentet fra Thises VMI-system. Regressionsanalysen bruges til at forudsige Coops efterspørgsel af koldskål samt andre variables effekt på denne.

Som supplement til de kvantitative data kombineres der med nogle kvalitative data i form af to semistrukturerede interviews med to informanter. Den ene informant arbejder i produktionsafdelingen, og den anden informant arbejder i salgsafdelingen. Formålet med den kvalitative data er, at den kan bruges til og forstå, hvorfor Thises datamodenhedsproces ser ud, som den gør på nuværende tidspunkt. Med

den kvalitative data kan vi derfor bevæge os dybere ned i vores problemstilling. Fremgangsmåde kaldes for metodetriangulering, idet vi vil kombinere tre forskellige datakilder til at besvare vores problemstilling med (ibid).

Den metodologiske fremgangsmåde betyder, at vi i større grad har arbejdet induktivt - dvs. hvor vi gik fra det konkrete til det generelle (ibid). Vi startede fx først ud med at lave en interviewguide og nogle spørgsmålsformuleringer, som vi operationaliserede ud fra vores hovedspørgsmål og vores underspørgsmål (se interviewguiden under bilag). Spørgsmålene blev stillet til vores informanter i to semistrukturerede interviews under vores besøg hos Thise Mejeri. Der har også været perioder, hvor vi har arbejdet deduktivt - dvs. hvor vi fik fra det generelle til det konkrete. Fx fandt vi på forhånd ud af nogle teorier om, hvorfor COOPs efterspørgsel på koldskål faldt eller steg. Det viste sig, at fx. at vejret og konkurrenternes pris på koldskål påvirkede efterspørgslen.

Den statistiske metode som anvendes i denne undersøgelse, kaldes for superviseret metode, fordi den tager udgangspunkt i en afhængig variabel. Den konkrete metode er en multibel lineær regression. Den vil vi anvende til og forudsige efterspørgslen på koldskål i liter fremadrettet ud fra effekten af nogle uafhængige vejr-variabler. Efter vi først har trænet vores model på træningsdata og fået beregnet de nødvendige koefficienter, får alle variablerne i ligningen smidt en hat på toppen - dette indikerer prædiktion (Hastie et.al 2021). Den generelle formel er:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 ... + \hat{\beta}_p x_p + \epsilon$$

$\hat{y}$ er den forudsagte værdi af Y og $\hat{f}$ er et estimat for f. $\hat{Y}$ er desuden den afhæn-

gige variabel efterspørgsel i liter. x_i er udvalgte uafhængige variabler. $\hat{y} = f(X)$ indeholder variation som vi kan reducere ved, at bruge den korrekte SL-metode til, at beregne f med. Men det vil aldrig være en fejlfri model vi ender ud med. Fordi estimatet ϵ er tilfældige fejl eller støj, man ikke kan gøre noget ved og den type fejl vil altid være til stede (Hastie et.al 2021).

3 Kvalitativ analyse

3.1 Thises forretningsmodel

Et business model canvas er et illustrativt og strategisk værktøj som bruges til, at forstå forskellige aspekter af en forretningsmodel (Ostervalder & Pigneur 2010). Det anvendes i denne sammenhæng til at forstå Thises forretningsmodel.

Thises værdiproposition er , at producere økologiske mælkeprodukter af høj kvalitet. Deres største kundesegment er COOP Danmark, og det er også dem som efterspørger størstedelen af koldskålen. De forhandler kun B2B, så de har ikke med private forbrugere at gøre. Derudover bruger de et VMI system som salgskanal. Et VMI system er et leverandørstyret system. COOP bestiller deres koldskål igennem dette system, og Thise producere dernæst koldskålen og opbevarer på deres eget lager. Dernæst transportere de selv koldskålen ud til COOPs butikker. Thise har en intern aftale med COOP om, at de skal have mindst 80% af af produkterne på lageret. De resterende 20% supplerer de op med vha. Pareto forecasting, som er et endnu et system i deres salgskanal. 80/20 reglen fra Pareto fungerer som

et management værktøj, da Thise hele tiden skal tilpasse deres salgsstrategier for at opretholde deres leveringsservice på 98 %.

Pareto er et godt værktøj til, at sikre at Thise opnår forskellige KPI målsætninger. Problemet med Pareto er, at det er abonnementbaseret og dyrt, fik vi fortalt af Bjarne som er senior demand planner (Justesen 2022: 59). Der er en stor risiko forbundet ved, at have et VMI system. Bliver et produkt ikke solgt, leveres det tilbage til Thises hovedlager som spild. Derfor er det vigtigt, at have nogle pålidelige forecasting modeller der kan forudsige efterspørgslen af produkter med stor præcision. Thise får nemlig først betaling når varen er solgt i COOPs butikker.

Det kræver et tæt samarbejde, at have et VMI system kørende. Derfor er det vigtigt at Thise overholder deres produktvilkår til COOP, da de er bundet af klausulaftaler som kan medføre forskellige sanktioner af økonomisk karakter, så frem de ikke overholdes. Stærke kunderelationer er derfor ekstremt vigtige i den her sammenhæng. I forbindelse med vores interview fortalte vores informant, at Thise har kunder som er meget loyale. Det betyder, at de ofte gerne vil gå lidt på kompromis med prisen (Jensen 2022: 59). Thises primære indkomststrøm er B2B, dog har de også en ostebutik. Dertil har de et samarbejde med Irma når de skal lancere nye og spændende produkter, bliver disse solgt i Irma butikker rundt i landet.

Thise samarbejder med landmændene som er deres nøgle partner, de bidrager økonomisk og kollektivt ift. at støtte op omkring økologisk landbrug og mejeri. Thises nøgleaktiviteter er, at de vil fremstille økologiske mælkeprodukter der fortæller den gode historie, og som på samme tid smager godt. Pris og kvalitet hænger sammen, og deres produkter skal ikke være de billigste på markedet, da de ikke vil gå på

kompromis med kvaliteten. Det er nødvendigt fordi Thise har store omkostninger i form lønninger, drifts - og produktionsomkostninger, afregning til landmanden og transport.

3.2 Thises datamodenhed

Alexandramodellen er en datamodenhedsmodel. Den bruges til at finde af hvor langt en virksomhed er i datamodenhedsprocessen, og hvordan den kan bruge forskellige strategier til og blive mere datadrevne (Bækby et. al 2017).

Alexandramodellen er opdelt i fem faser til forandring af dataanvendelse. Første fase er at opsamle og monitorere drift og forandring (ibid). Thise Mejeri er på nuværende tidspunkt i denne fase, da de registrerer data for at skabe en forståelse og værdi for virksomhedens drift (ibid). Thise Mejeri opsamler data i form af produktion og salg fra tidligere perioder, og dette kan dermed bruges til at analysere og forudsige fremtidige produktion. Dataindsamlingen hertil har ikke en stor omkostning eller processeringstid, da flere af produktionsmaskinerne kan opsamle informationer om dette løbende. Dertil bliver dataen behandlet i Pareto af en udbyder som forecaster dataen og indsender det til Bjarne, som er demand planner hos Thise Mejeri. Bjarne undersøger dataen igennem inden det bliver sendt videre til de forskellige afdelinger.

Salget fra kunderne bliver indsamlet i Navision. Bjarne fortæller i interviewet, at Thise Mejeri er meget fokuseret på ordrehistorikken fra tidligere salg for at kunne forecaste produktionen. Alexandramodellens første fase skaber desuden en stor værdi for Thise Mejeri, da det giver medarbejderne en hjælpende hånd til at

forudsige produktionen for næste periode. Flere afdelinger gør godt brug af dataindsamlingen, og analysen kan bruges til produktions-, marketing- eller i salgsafdelingen.

Selvom første fase skaber en stor værdi, kan det også skabe en kulturel udfordring. Søren fra produktionsafdelingen fortæller i interviewet, at medarbejderne hos Thise Mejeri har en generel høj anciennitet og har dermed være i virksomheden i mange år. Dette kan skabe en udfordring i forhold til forandring. Han nævner også at it-kompetencerne i produktionen ikke er høj, og det kan derfor skabe udfordringer i forbindelse med at integrere medarbejderne for dataindsamlingen. Det er derfor vigtigt at have en gennemsigtighed i denne fase, for at medarbejderne har en forståelse for dette og dataindsamlingens vigtighed for fremtidig produktion.

Produktionsplanlæggeren har via. Navision adgang til alt virksomhedsdata gennem deres Windows styresystem, hvilket er en on-premises-løsning. Navision er ERP software til at håndtere økonomi data til og træffe beslutninger ud fra. og kan dermed planlægge fremtidig produktionen. At de anvender programmet direkte gennem deres Windows styresystem gør virksomheden sårbar overfor systemnedbrud, hvilket de oplever ca 5 gange om året. Kompetencemæssigt, ifølge Bjarne, har Thise Mejeri en brist da de ikke har ansat nogle medarbejder med relevant analyse erfaring. De er derfor afhængige af den eksterne udbydere til at forecaste med de data, der bliver indsamlet.

4 Introduktion til dataanalysen

Thise har svært at forudsige præcist hvor mange liter koldskål produktionsafdelingen skal producere. Dette har resulteret i, at de ikke har kunne producere nok koldskål i år, fordi flere af butikkerne i området har oplevet at deres køldiske har været tomme for koldskål i sommerperioden.

4.1 Baggrund

Den kvantitative del af analysen er afgrænset til COOP butikker i nærheden af Landbohøjskolen, hvis beliggenhed er i Københavnområdet. Butikkerne afgiver ordre til Thises fjernlager, hvorefter koldskålen bliver leveret ud til butikkerne.

4.2 Formål

Formålet med analysen er derfor, at udregne en multibel lineær regressionsmodel som bedst kan forudsige butikkernes efterspørgsel på koldskål i området omkring Landbohøjskolen. Derudover vil vi også finde ud af, hvordan vejret og andre vejrrelateret faktorer påvirker butikkernes efterspørgsel på koldskål. Denne fremgangsmåde kan løse Thises forretningsproblem. Undersøgelsens dataminingproblem går ud på, at identificere de forskellige vejr-variablers effekt på efterspørgspørgslen af koldskål.

```
pacman::p_load("tidyverse", "magrittr", "nycflights13", "gapminder",
               "Lahman", "maps", "lubridate", "pryr", "hms", "hexbin",
               "feather", "htmlwidgets", "broom", "pander", "modelr",
               "XML", "httr", "jsonlite", "lubridate", "microbenchmark",
               "splines", "ISLR2", "MASS", "testthat", "leaps", "caret",
               "RSQLite", "class", "babynames", "nasaweather",
               "fueleconomy", "viridis", "readxl", "timeDate","tinytex",
               "ggbeeswarm", "palmerpenguins", "hms", "RColorBrewer",
               "boot", "openxlsx", "writexl", "PerformanceAnalytics",
               "car", "pscl", "caret")
```

4.2.1 Importer data til R

I første omgang importeres datasættet:

```
# Indlæser datasæt og gemmer det nye datasæt i et objekt.
data1 <- read_excel("data/stud_exam_data.xlsx")
 # Dernæst undersøges strukturen i datasættet.
#str(data1)
```

4.3 Tidying og transformering af datasæt

Nu har vi fået indlæst datasættet. Det næste skridt er gøre strukturen i vores dataframe nemmere at arbejde med og mere læsevenlig. Denne proces kaldes for

tidy data, det betyder at hver variabel har en kolonne, hver observation en række, samt hver observationsenhed er i en tabel (Wickham 2022). Det gør analysearbejdet nemmere. Først rekode nogle af variablerne, så de stemmer overens med hvad der står i opgavebesvarelsen.

I nedestående kodestump rekodes og transformeres de udvalgte variabler så de stemmer overens med eksamensbesvarelsen. Hele kodestumpen vil blive kædet sammen med 'pipe' funktionen' fra dplyr pakken. Omkodningerne bliver til sidst gemt i en ny dataframe som kaldes data1.

Derefter bruges `mutate()` til at lave en ny kolonne ud fra data1. Først laves der en date-variabel, som bliver kodet om til et date objekt med `ymd()` fra lubridate pakken.

I den næste del anvendes mutate igen til, at lave en kolonne der hedder dag, som bliver omkodet til en faktor. Dernæst koder vi date til et objekt med `ymd()` funktionen fra lubridate pakken. "lubridate.week.start",1=mandag, istedet for søndag som er standardindstillingerne i R.

Dernæst bruger vi `mutate()` igen til at danne en ny weekend-variabel der hedder weekend_1. I denne sammenhæng vælger vi at fredag, lørdag, søndag og fire andre helligdage er 1, ellers er de andre værdier 0. Dette kaldes for en dummyvariabel.

Måned, dag, kamjunk, forvent_lager og weekend_1 er alle kategoriske faktorer. For at gøre det nemmere at forstå hvad de forskellige værdier udtrykker, navngives disse med `fct_recode()` funktionen.

4.3.1 Den Afhængige variabel

Efterspørgsel er den afhænge variabel på et ratiointervalskaleret måleniveau, fordi den har et naturligt nulpunkt. Dertil kan man beregne afstanden fra 100 liter koldskål til 150 liter koldskål. Enheden er liter.

4.3.2 Uafhængige variabler

Der er seks uafhængige variable på nominalt måleniveau, de er kodet som faktorer med forskellige kategorier. Antagelsen er at de alle sammen har en effekt på den afhængige variabel efterspørgsel på koldskål i liter.

```
data1 <- read_excel("data/stud_exam_data.xlsx") %>%
  mutate(date = ymd(date), måned = factor(month(date)),
  kamjunk = factor(kammerjunkere), forvent_lager =
  factor(forventet_l_lager)) %>%
  mutate(dag = as.factor(wday(date, week_start =
  getOption("lubridate.week.start", 1)))) %>%
  mutate(weekend_1 = as.integer(dag %in% c("5", "6", "7")| date %in%
  ymd("2022-04-14", "2022-04-18", "2022-05-26","2022-06-06"))) %>%
  mutate(weekend = factor(weekend_1)) %>%
  mutate(data1, kamjunk = fct_recode(kammerjunkere,
                                     "ja" = "0",
                                     "nej" = "1")) %>%
  mutate(data1, forvent_lager = fct_recode(forventet_l_lager,
```

```
                                        "lav" = "1",
                                        "mellem" = "2", "høj" = "3")) %>%
  mutate(data1, måned = fct_recode(måned, "april" = "4", "maj" = "5",
                                        "juni" = "6", "juli" = "7",
                                        "august" = "8")) %>%
  mutate(data1, dag = fct_recode(dag, "mandag" = "1", "tirsdag" = "2",
                                        "onsdag" = "3", "torsdag" = "4",
                                        "fredag" = "5", "lørdag" = "6",
                                        "søndag" = "7")) %>%
  dplyr::select(date, måned, dag, efterspørgsel, kamjunk, forvent_lager,
  weekend_helligdag = weekend)
glimpse(data1)
```

```
Rows: 152
Columns: 7
$ date              <dttm> 2022-04-01, 2022-04-02, 2022-04-03, 2022-04-04, 2
$ måned             <fct> april, april, april, april, april, april, april, a
$ dag               <fct> fredag, lørdag, søndag, mandag, tirsdag, onsdag, t
$ efterspørgsel     <dbl> 367, 361, 376, 47, 367, 402, 416, 355, 283, 454,
$ kamjunk           <fct> ja, ja, ja, nej, ja, ja, ja, ja, nej, ja, nej, ja
$ forvent_lager     <fct> høj, høj, høj, høj, høj, høj, høj, høj, høj, høj,
$ weekend_helligdag <fct> 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1
```

I denne kodechunk vil vi lave en HTTP GET-anmodning til en API fra DMI. Vi skal bruge adgangen til at få de relevante vejr-variable som vi senere skal

bruge i vores analyse. API'en leverer til slut et objekt i JSON format som bliver transformeret om til en dataframe i stedet for en liste.

Først bruger vi base_url og info_url til at anmode om vejrdata fra DMI's API. req_url bruges til at udvælge specifikke parametre fra API´en.

I denne kodechunk vil vi transformere den data vi har hentet fra vores API-kald til nogle mere brugbare data.

Først bruger vi base_url og info_url til at anmode om vejrdata fra DMI's API. req_url bruges til at udvælge specifikke parametre fra API´en.

Derefter bruges `pivot_wider()` til at fordele variablerne ud i deres egne separate kolonner.

Vi bruger derefter mutate-funktionen til at konvertere kolonnen målingstidspunkt til datoformat. `Separate()` bruges til at opdele kolonnen målingstidspunkt i to separate kolonner som vi navngiver 'date' og 'time'.

`Filter(str_sub)`funktionen udvælger rækker, der indeholder de første fire karakterer: "12:0".

I nedestående kode-chunk bruges `left_join()` til at returnere alle rækkerne fra x, samt alle kolonner langs x og y (Wickham 2022). Udvalgte dataframes bliver sammenkoblet fra data1 og data2 til data3. Da alle kolonnerne er lige lange, optræder der ingen missing værdier.

Dernæst anvendes mutate til, at oprette fire nye variabler i data3 kaldet temp_gt25_3_dage. `Lag()`-funktionen er brugt til at lave variablerne, som har opfanget forsinkede værdier fra temp1, temp2 og temp3.

Afslutningsvis dannes variablen 'temp_gt25_3_dage', som måler de dage hvor der har været mere end 3 dage i træk med >= 25 grader. Det er en dummyvariabel fordi der bruges `if_else`. Relevante variabler bliver beskrevet når der tolkes på modelparametre i regressionsanalysen.

```
data3 <- data1 %>%
left_join(data2, data1, by = c("date" = "date"))
#dplyr::select(data3, date, time, weekend_helligdag, everything())



data3 <- data3 %>%
  mutate(temp1 = lag(temp_max_past1h, 1),
         temp2 = lag(temp_max_past1h, 2),
         temp3 = lag(temp_max_past1h, 3),
         temp1 = if_else(is.na(temp1), 0, temp1),
         temp2 = if_else(is.na(temp2), 0, temp2),
         temp3 = if_else(is.na(temp3), 0, temp3),
         temp_gt25_3_dage = if_else(temp1 >= 25 & temp2 >= 25 & temp3 >= 
```

4.4 Datavisualisering og eksplorativ analyse

Nu er data blevet gjort tidy. Næste skridt er at undersøge hvilke vejr-mønstre der hænger sammen med butikkernes efterspørgslen af koldskål i det sammenkoblede datasæt, som vi har kaldt data3. I næste afsnit starter den eksplorative analyse.

Først identificeres potentielle outliers i vores dataset. Der fjernes 1 outlier som er 47, fordi den skiller sig væsentligt ud i forhold til de andre observationer. Måske denne er en tastefejl. Umiddelbart vurderes det ikke, at der er mange outliers i data som kan have indflydelse på den samlede varians, hvorfor der kun er fjernet den ene. Tilbage er der n = 151 i data3.

Derefter laves der et ggplot for at se fordelingen af efterspørgselen af koldskål i form af simpelt histogram, fordi efterspørgslen er en kontinuert variabel. Det er derfor muligt, at beregne spredningen mellem observationerne.

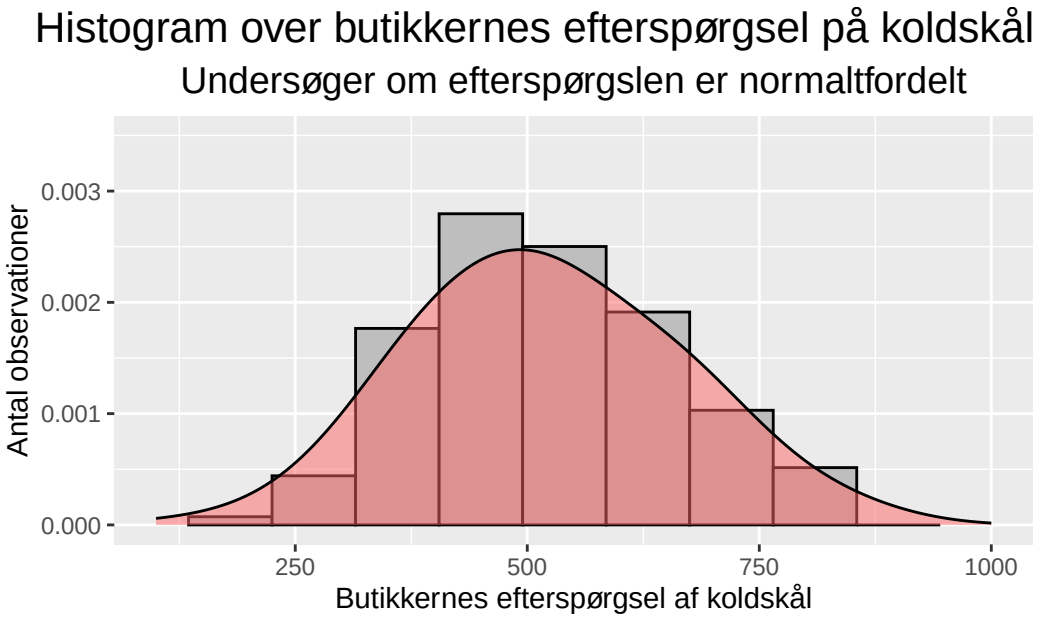
`geom_density()` funktionen bruges til, at forstå fordelingen og til at forudsige den forventede fordeling af efterspørgslen på koldskål. Man kan se at at spredningen af observationerne er størst omkring 520 liter. Endvidere kan det ses, at efterspørgslen af koldskål er tilnærmelsesvis normalfordelt, og at sandsynlighedskurven er symmetrisk klokkeformet.

Dog kan man også se, at nogle af observationerne falder udenfor, hvilket kan skyldes tilfældig variation eller systematiske fejl. Man ved desuden, at ca. 50% af observationerne befinder sig til venstre og højre af midtpunktet, dette er middelværdien.

At data er normalfordelt er en fordel, fordi den lineære regressionsmodel er en parametrisk test, hvor ét af kravene er at data er skal være normaltfordelt (Hastie et.al 2021). At dette krav er opfyldt gør envidere, at regressionsmodellens parametre er mere pålidelige.

```r
data3 <- data3 %>%
  filter(efterspørgsel > 47) # Fjerner obsnr. 47 fra datasættet.

ggplot(data3, aes(x = efterspørgsel)) +
 geom_histogram(aes(y = ..density..), color = "black",
                fill = "grey", binwidth = 90) +
 geom_density(alpha = 0.5, fill="#FF6666", adjust = 1.6) +
  labs(title = "Histogram over butikkernes efterspørgsel på koldskål i lt
  subtitle = "Undersøger om efterspørgslen er normaltfordelt",
  y = "Antal observationer",
  x = "Butikkernes efterspørgsel af koldskål",
  caption = "Kilde: Tal fra DMI 2002. Fra perioden 1/4/22-30/8/22") +
  theme( plot.title = element_text(hjust = 0.5, size = 16),
         plot.subtitle = element_text(hjust = 0.5, size = 14),
         plot.caption = element_text(hjust = 1, face = "italic", size = 1
  xlim(100, 1000) + ylim(0, 0.0035)
```

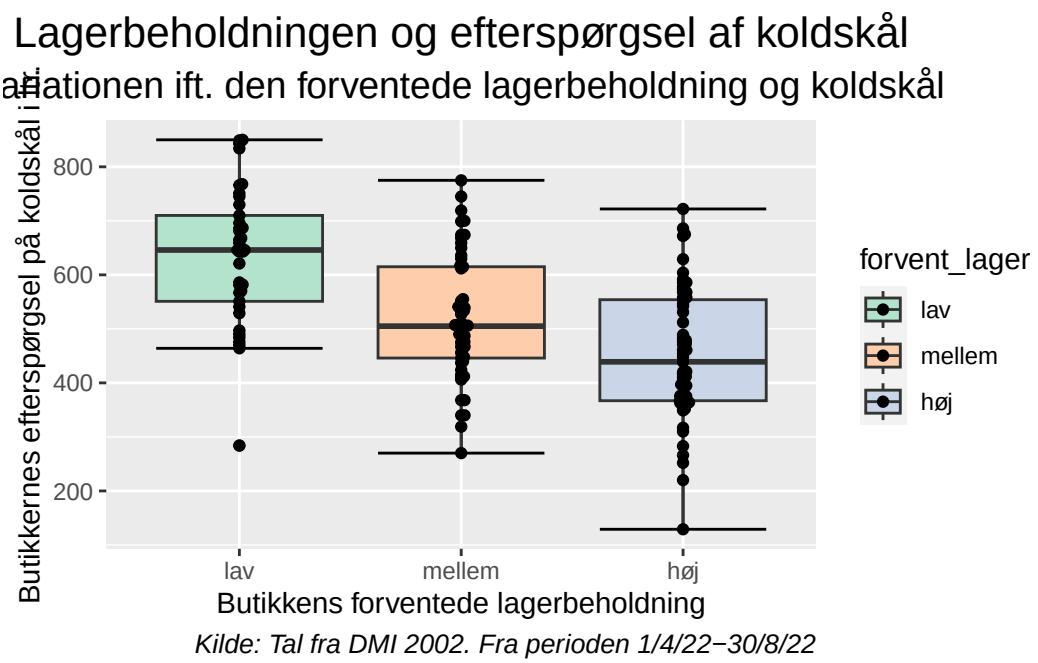


I følgende kode-chunk har er der lavet et boxplot. Det skal vise den statistiske variationen ift. butikkens forventede lagerbeholdning og efterspørgslen på koldskål.

Her kan man se at median-efterspørgslen stiger når man går fra høj til lav forventet lagerbeholdning af koldskål. Dette tyder også på at der er en signifikant sammenhæng mellem de 2 variabler. Man kan også vha. `geom = 'errorbar'` se, at der er fx. ved en høj forventet lagerbeholdning er en relativ stor usikkerhed i forhold til mellem og lav lagerbeholdning, det indikerer at datapunkterne er forholdsvis meget spredt ud.

```
ggplot(data = data3, mapping = aes(x = forvent_lager,
                                   y = efterspørgsel,
                                   fill = forvent_lager)) +
  stat_boxplot(geom = 'errorbar') + # Viser usikkerheden.
  geom_boxplot() +
```

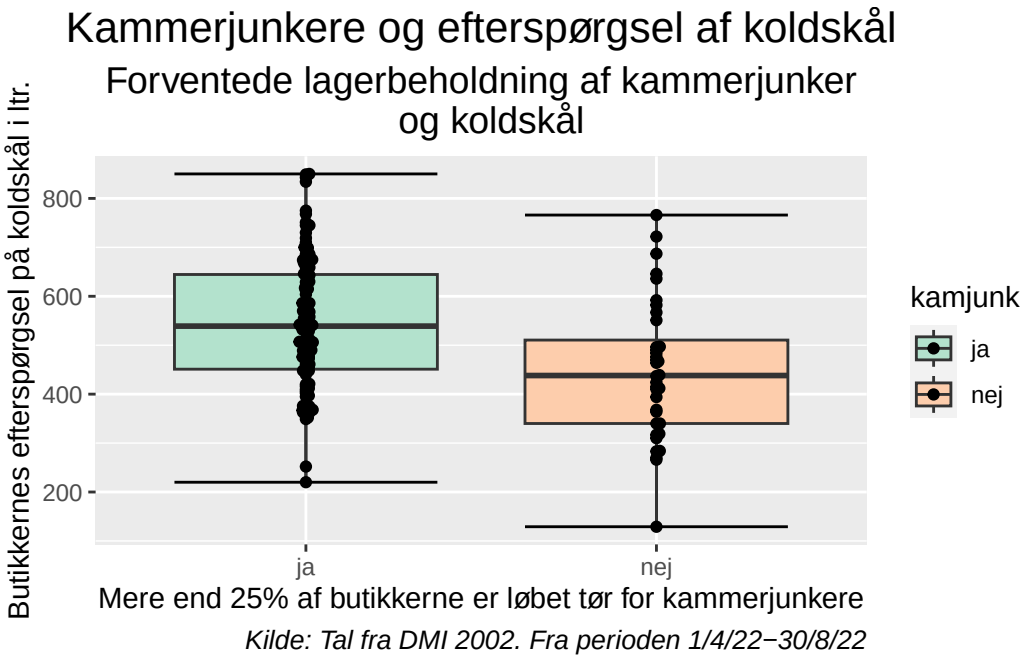
```
  labs(title = "Lagerbeholdningen og efterspørgsel af koldskål",
    subtitle = "Variationen ift. den forventede lagerbeholdning og koldska
    caption = "Kilde: Tal fra DMI 2002. Fra perioden 1/4/22-30/8/22",
y = "Butikkernes efterspørgsel på koldskål i ltr.",
x = "Butikkens forventede lagerbeholdning") +
  geom_beeswarm(dodge.width=0, cex = 0.5, color = "black") +
  theme( plot.title = element_text(hjust = 0.5, size = 16),
         plot.subtitle = element_text(hjust = 0.5, size = 14),
         plot.caption = element_text(hjust = 1, face = "italic",
                                     size = 10 )) +
scale_fill_brewer(palette = "Pastel2")
```



I næste kode-chunk er der lavet et boxplot som viser fordelingen af efterspørgslen i forhold til om 25% af butikkerne er løbet tør for kammerjunkere eller ej.

På baggrund af plottet kan man se at hvis butikkerne ikke har kammerjunkere på lageret så falder efterspørgslen. Det betyder at Efterspørgslen på koldskål stiger hvis butikkerne er løbet tør for kammerjunkere.

```
ggplot(data = data3, mapping = aes(x = kamjunk, y = efterspørgsel, fill =
                                     kamjunk)) +
  stat_boxplot(geom = 'errorbar') +
  geom_boxplot() +
  labs(title = "Kammerjunkere og efterspørgsel af koldskål",
  subtitle = "Forventede lagerbeholdning af kammerjunker
  og koldskål",
  caption = "Kilde: Tal fra DMI 2002. Fra perioden 1/4/22-30/8/22",
  y = "Butikkernes efterspørgsel på koldskål i ltr.",
  x = "Mere end 25% af butikkerne er løbet tør for kammerjunkere") +
  geom_beeswarm(dodge.width = 0,cex = 0.5, color = "black") + # Justerer
  theme( plot.title = element_text(hjust = 0.5, size = 16),
         plot.subtitle = element_text(hjust = 0.5, size = 14),
         plot.caption = element_text(hjust = 1, face = "italic",
                                     size = 10 )) +
scale_fill_brewer(palette = "Pastel2")
```

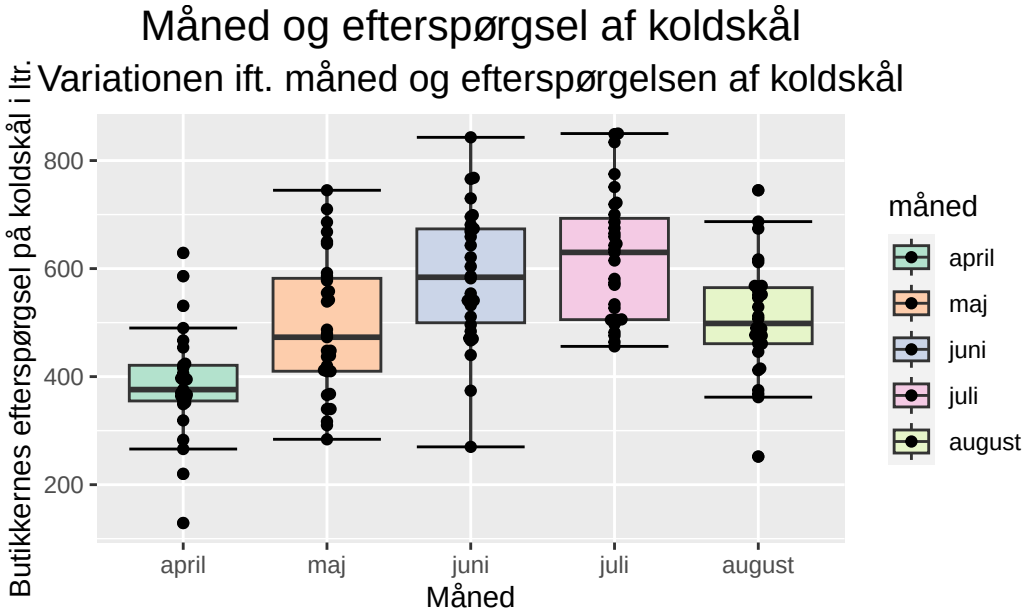


Forneden er et boxplot som viser sammenhængen mellem måned og efterspørgslen af koldskål. Det er tydeligt at se at median-efterspørgslen stiger fra april-juli hvorefter den efterspørgslen igen falder i august.

Den overordnede observation stemmer også overens med påstande fra vores kilder i problemfeltet, og udsagn fra vores interviews med medarbejderne hos Thise Mejeri. Dette indikerer at efterspørgselen på koldskål hænger moderat sammen med årstiden, dvs. selve sommerperioden, da de to boxplots ikke overlapper hinanden. I næste afsnit undersøges sammenhængen mere dybdegående.

```
ggplot(data = data3, mapping = aes(x = måned,
                                   y = efterspørgsel,
                                   fill = måned)) +
  stat_boxplot(geom = 'errorbar') +
  geom_boxplot() +
```

```
  labs(title = "Måned og efterspørgsel af koldskål",
  subtitle = "Variationen ift. måned og efterspørgelsen af koldskål",
  caption = "Kilde: Tal fra DMI 2002. Fra perioden 1/4/22-30/8/22",
  y = "Butikkernes efterspørgsel på koldskål i ltr.",
  x = "Måned") +
  ggeasy::easy_center_title() + # Centrerer titlen.
  geom_beeswarm(dodge.width = 0,cex = 0.5, color = "black") + # Justerer
  theme(plot.title = element_text(hjust = 0.5, size = 16),
        plot.subtitle = element_text(hjust = 0.5, size = 14),
        plot.caption = element_text(hjust = 1, face =
                                      "italic", size = 10 )) +
scale_fill_brewer(palette = "Pastel2")
```



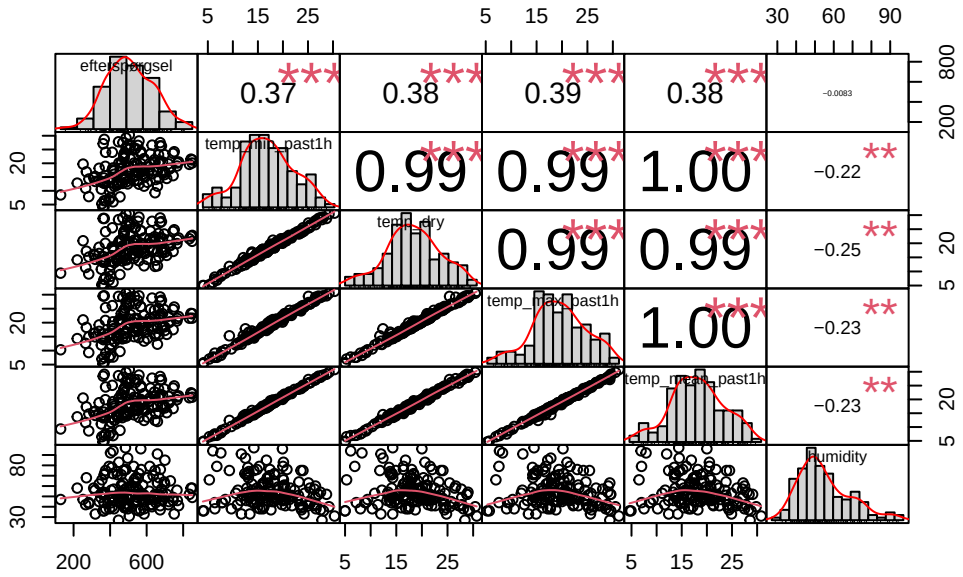
I nedestående kodechunk er der udvalgt 6 kontinuerte variabler, fordi ønskes er, at

undersøge om disse korrelerer med hinanden, og om deres indbyrdes korrelation er statistisk signifikant. Der anvendes en `chart.Correlation()` til at foretage en korrelationsanalyse.

Efterspørgel og humidity er ikke korreleret med hinanden, og dermed ikke statistisk signifikant. Beslutningen er derfor, at humidity ikke vil blive inkluderet i analysen. Efterspørgsel og den gennemsnitlige temperatur per time har en korrelations koefficient på 0.38. P-værdien er meget lav med tre stjerner, det betyder at sammenhængen er signifikant, og det er derfor usandsynligt at opnå et mere ekstremt resultat, hvis man foretog en ny stikprøve igen og igen.

Alle temperatur-variablerne er tæt på 1, hvilket betyder at de har stærk samvariation. Dette kaldes for multikolinearitet. Det vil sige, hvis de blev brugt i den endelige model ville det være vanskeligt, at fortolke på koefficienterne. Fordi en Model med høj multikolinearitet bliver mindre præcis og mindre pålidelig. For at reducere multikolineariteten fjernes de øvrige temperatur-variable fra modellen.

```
cor_matrice <- data3 %>%
  dplyr::select(efterspørgsel,
                temp_min_past1h,
                temp_dry,
                temp_max_past1h,
                humidity)
chart.Correlation(cor_matrice, histogram = TRUE, method = "pearson")
```



Som førnævnt var der en moderat signifikant sammenhæng mellem efterspørgslen og gennemsnits temperaturen. Derfor bruges temp_mean_past1h som den uafhængige effekt i næste kodechunk. Der anvendes `predict()` til at konstruere et 95 prædiktionsinterval, efterfulgt af `geom_smooth()` til og visualisere sammenhængen med et scatterplot.

Ud fra scatterplottet kan man se, at forholdet mellem den gennemsnitlige temperatur og butikkernes efterspørgsel på koldskål er moderat lineært. Fordi hældningen på tendenslinjen er positiv. Vi antager at når gennemsnits temperaturen stiger én enhed, vil efterspørgslen stige tilsvarende, da mange af datapunkterne er tæt på linjen. Der anvendes lineær regression, fordi det er en simpel metode, hvor det er nemt og tolke på parametrene.

Men mange af datapunkterne ligger også langt væk fra tendenslinjes som udtrykker en stigende gennemsnitlig efterspørgsel på koldskål i liter. Dette indikerer at der er stor varians og bias. En mere kompleks model skal derfor anvendes til, at forklare

sammenhængen.

Efterspørgslen af koldskål bliver prædiktet ud fra den gennemsnitlige temperatur hver time i °C. Først ved 10, 20 og 30 grader. Den grå linje omkring tendenslinjen referer til konfidensintervallet af den gennemsnitlige efterspørgsel på koldskål ved en given temperatur.

Mange af observationerne er placeret udenfor dette bånd, hvorfor det er besluttet at anvende et prædiktionsinterval i stedet - der er den røde stiplede linje. Formålet er med andre ord, at indfange usikkerheden omkring de individuelle værdier og ikke usikkerheden omkring gennemsnittet.

Når den gennemsnitlige temperatur hver time er 10 °C, forudsiger vi at efterspørgslen på koldskål for én ny observation vil være 440.64 liter. Ved samme temperatur vil efterspørgslen med 95% sikkerhed være [181.78:699.51].

Når den gennemsnitlige temperatur hver time er 20 °C, forudsiger vi at efterspørgslen på koldskål for én ny observation vil være 535.25 liter. Ved samme temperatur vil butikkernes efterspørgslen af koldskål med 95% sikkerhed være [278.23:792.28].

Når den gennemsnitlige temperatur hver time er 30 °C, forudsiger vi at efterspørgslen på koldskål for én ny observation vil være 629.87 liter. Ved samme temperatur vil butikkernes efterspørgslen af koldskål med 95% sikkerhed være [369.30:890.43].

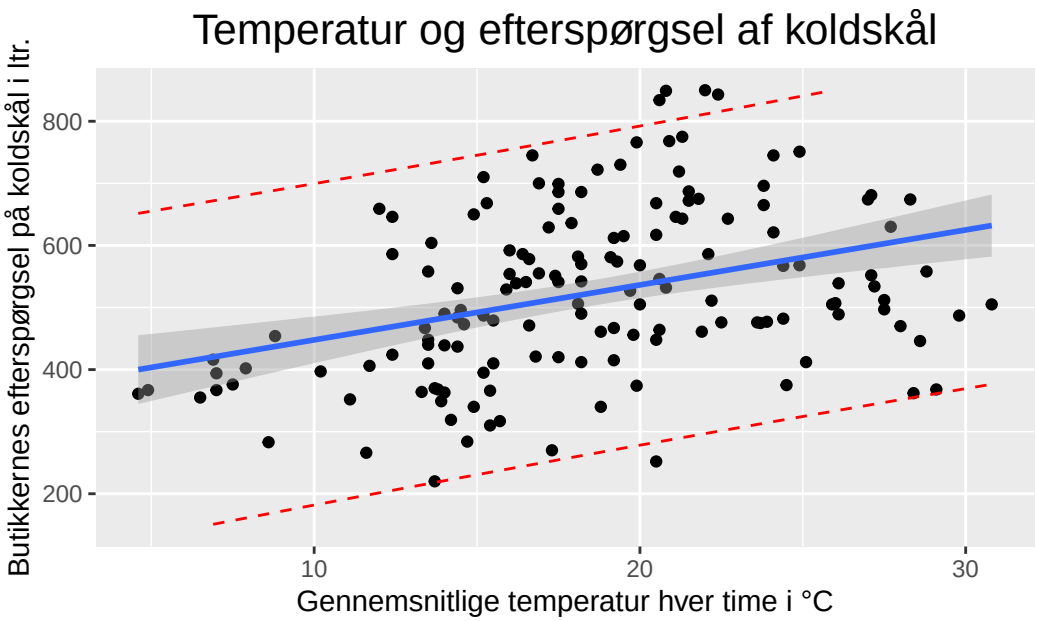
Man kan på baggrund af ovenstående tydeligt se, at hvis den gennemsnitlige temperatur i °C stiger, så stiger butikkernes efterspørgsel på koldskål tilsvarende.

```r
model1 <- lm(efterspørgsel ~ temp_mean_past1h, data = data3)
predict(model1,data.frame(temp_mean_past1h = (c(10,20,30))),
        interval = "prediction", level = 0.95)
```

```
      fit      lwr      upr
1 440.6455 181.7817 699.5094
2 535.2564 278.2290 792.2838
3 629.8672 369.3044 890.4301
```

```r
prædiktion <- predict(model1, interval = "prediction", level = 0.95)
ny_df <- cbind(data3, prædiktion)
ggplot(ny_df, aes(temp_mean_past1h, efterspørgsel)) +
    geom_point() +
    geom_line(aes(y=lwr), color = "red", linetype = "dashed") +
    geom_line(aes(y=upr), color = "red", linetype = "dashed") +
    geom_smooth(method = lm, se = TRUE) +
    labs(title = "Temperatur og efterspørgsel af koldskål",
    caption = "Kilde: tal fra DMI 2002 fra perioden 1/4/22-30/8/22",
    y = "Butikkernes efterspørgsel på koldskål i ltr.",
    x = "Gennemsnitlige temperatur hver time i °C") +
ggeasy::easy_center_title() + # Centrerer titlen.
theme( plot.title = element_text(hjust = 0.5, size = 16),
plot.subtitle = element_text(hjust = 0.5, size = 10),
plot.caption = element_text(hjust = 1, face = "italic", size = 10 )) +
```

```
xlim(4.6, 30.8) + ylim(150, 850)
```



Kilde: tal fra DMI 2002 fra perioden 1/4/22−30/8/22

4.5 Træning på træningsdata

I første omgang trænes modellen på træningsdata, fordi vi gerne vil tilpasse modelparametrene. Vi bruger træningsdata til, at fintune vores regressionsmodel. Når modellen er blevet trænet godt igennem, bliver den afprøvet på testdata, da vi gerne vil undersøge hvor god modellen er til, at forudsige en så præcis efterspørgsel på koldskål som mulig. Vurderingen af modelpræcisionen bestemmes ud fra den laveste MSE værdi. MSE måler hvor langt den forudsagte værdi for en observation er fra den faktiske værdi for en observation.

Er MSE lille er der den forudsagte værdi tæt på den faktiske værdi, er MSE stor er den forudsagte værdi langt fra den faktiske værdi. MSE er således et udtryk

for, hvor præcis den udvalgte model er til, at forudsige efterspørgslen af koldskål (Hastie et.al 2021).

Da antallet af variabler i det samlede datasæt er mindre end antallet af observationer bruges backward-selection til, at udvælge de uafhængige variabler som fremadrettet skal indgå i modellerne. Det vil sige, at vi tilføjer alle variable ind på højre side af ligningen, og fjerner dem med den højeste p-værdi indtil der kun er signifikante uafhængige variable tilbage (Hastie et.al 2021).

Denne teknik kan hjælpe med at reducere unødvendig varians i den udvalgte model, men på samme tid være effektiv nok til at identificere vigtige relationer i datasættet (ibid).

```
# Baseline model


lm.fit_træning <- lm(efterspørgsel ~ 1, data = data3)
lm_fit_summary <- summary(lm.fit_træning)
#mean(lm_fit_summary$residuals^2) # MSE 19376.55
rmse(lm.fit_træning, data = data3)  # RMSE = 139.19
```

```
[1] 139.1997
```

```
#lm_fit1.1_summary


# Simpel model


lm.fit2_træning <- lm(efterspørgsel ~ temp_mean_past1h, data = data3)
```

```r
lm_fit2_summary <- summary(lm.fit2_træning)
#mean(lm_fit2_summary$residuals^2) #  MSE = 16576.45
rmse(lm.fit2_træning, data = data3)  # RMSE = 128.74
```

```
[1] 128.7495
```

```r
#lm_fit2_summary

# Mellem model

lm.fit3_træning = lm(efterspørgsel ~ forvent_lager +
                       weekend_helligdag + kamjunk +
                       temp_gt25_3_dage + måned +
                       I(temp_mean_past1h), data = data3)
lm_fit3_summary <- summary(lm.fit3_træning)
#mean(lm_fit3_summary$residuals^2) # MSE = 6740.342
rmse(lm.fit3_træning, data = data3) # RMSE = 82.09959
```

```
[1] 82.09959
```

```r
#lm_fit3_summary

# Kompleks model

lm.fit4_træning = lm(efterspørgsel ~ forvent_lager +
```

```
                        weekend_helligdag +
                        kamjunk + temp_gt25_3_dage +
                        måned +
                        I(temp_mean_past1h^22), data = data3)
lm_fit4_summary <- summary(lm.fit4_træning)
#mean(lm_fit4_summary$residuals^2) # MSE = 6923.087
rmse(lm.fit4_træning, data = data3) # RMSE = 83.20509
```

```
[1] 83.20509
```

```
#lm_fit4_summary
```

4.6 Test på testdata

I fornævnte afsnit blev den gennemsnitlige MSE beregnet for hver af de fire modeller på træningsdata. Vi er egentlig ligeglade med disse MSE værdier, det er mere interessant at se hvor præcise forudsigelserne er på testdata. Træningsdata anvendes som førnævnt til, at udvælge signifikante uafhængige variable og tilpasse modelparametrene.

Dog er det værd at nævne, at den data undersøgelsen er baseret på simulerede data. Dvs. at f er kendt allerede. Den virkelige og komplekse sandhed om efterspørgslen af koldskål vides dog ikke. Men hvis der bliver udtrukket nogle testdata ud fra

data3, kan man validere hvor godt en model performer på disse testdata, når modelkompleksiteten øges. Kompleksiteten kan øges ved, at de kontinuerte variable opløftes i flere potenser, eller ved og inkludere flere uafhængige variabler.

4.7 Valg af metode til at teste model performance

Man kan producere testdata på flere måder. Der anvendes i denne forbindelse en LOOCV metode, fordi data3 kun indeholder 151 observationer i alt. Fordelen ved denne fremgangsmåde er, at den træner på alle observationerne, undtagen ét datapunkt. Processen gentages i dette tilfælde 150 gange. Derefter beregnes en gennemsnitlig MSE score, som udtrykker hvor god modelpræcisionen er (Hastie et.al 2021).

Problemet med metoden er, at det kræver stor computerkraft. Det tog denne bærbare computer ca. 2 minutter hver gang kodestumpen blev kørt. Det skyldes at modellen trænes k gange (ibid).

4.8 Fortolkning af multible lineær regression

Med udgangspunkt i tabel 1 tolkes der på modelparametrene.

4.9 Baseline model

4.10 Simpel model

4.11 Mellem model

4.12 Kompleks model

Vurdering af outliers

	Baseline	Simpel	Mellem	Kompleks
Forvent_lager(mellem)			−76.57***	−74.55***
			{19.82}	{19.72}
Forvent_lager(høj)			−82.67***	−87.00***
			{22.58}	{22.81}
Weekend_helligdag(ja)			113.01***	107.33***
			{15.00}	{15.16}
Kamjunk(nej)			−71.89***	−76.52***
			{17.50}	{19.82}
Temp_gt25_3_dage			−82.00*	−66.74*
			{34.23}	{34.90}
Måned(maj)			84.37***	101.69***
			{24.54}	{23.36}
Måned(juni)			129.51***	162.71***
			{32.79}	{27.87}

	Baseline	Simpel	Mellem	Kompleks
Måned(juli)			155.95***	155.947***
			{32.79}	{29.12}
Måned(august)			85.10*	197.44*
			{34.76}	{30.96}
Temp_mean_past1h		9.46***	4.249*	0.0
		{1.89}	{2.07}	
Uafhængige variable	0	1	6	6
Skæring $\hat{\beta}_0$	520.23***	346.04***	379.93***	434.78***
Model P-værdi	* * *	* * *	* * *	* * *
MSE_træning	139.20	128.74	82.10	83.21
MSE_test	139.20	130.36	88.55	89.37
R^2		0.15	0.65	0.64
Obs.	150	150	150	150

Tabel 1. Summeret modelreferat fra testdata. Referencegrupper () for faktorerne er: kamjunkja, forvent_lagerlav, månedapril. {} referer til standardfejlen. Note:$* = P < 0.1; ** = P < 0.05; *** = P < 0.01$

```
ctrl <- trainControl(method = "LOOCV") # Udvælger cross-validation metode

# Baseline model


lm.fit_træning <- glm(efterspørgsel ~ 1, data = data3)
```

```r
cv.err1 <- cv.glm(data3, lm.fit_træning)
cv.err1$delta[[1]]
```

```
[1] 19635.76
```

```r
rmse(lm.fit_træning, data = data3) # 139.1997
```

```
[1] 139.1997
```

```r
#lm.fit_træning <- lm(efterspørgsel ~ 1, data = data3) # Baseline model
#lm_fit1.1_summary <- summary(lm.fit_træning) # RMSE = 139.19
#summary(lm.fit_træning)

# Simpel model

model1_test <- train(efterspørgsel ~ temp_mean_past1h, data = data3,
                     method = "lm", trControl = ctrl)
model1_test # RMSE 130.36
```

```
Linear Regression

151 samples
  1 predictor
```

```
No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 150, 150, 150, 150, 150, 150, ...
Resampling results:

  RMSE      Rsquared   MAE
  130.3607  0.1238588  105.8906

Tuning parameter 'intercept' was held constant at a value of TRUE
```

```r
#summary(model1_test)

# Mellem model

model2_test <- train(efterspørgsel ~ forvent_lager +
                       weekend_helligdag + måned +
                       kamjunk +
                       temp_gt25_3_dage +
                       I(temp_mean_past1h^1), data = data3,
                       method = "lm", trControl = ctrl)
model2_test # RMSE 88.55
```

```
Linear Regression

151 samples
```

```
  6 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 150, 150, 150, 150, 150, 150, ...
Resampling results:

  RMSE      Rsquared   MAE
  88.55084  0.5967607  72.56083

Tuning parameter 'intercept' was held constant at a value of TRUE
```

```r
#summary(model2_test)

# Kompleks model

model3_test <- train(efterspørgsel ~ forvent_lager +
                       weekend_helligdag +
                       kamjunk +
                       temp_gt25_3_dage +
                       måned +
                       I(temp_mean_past1h^22), data = data3,
                       method = "lm", trControl = ctrl)
model3_test # RMSE 89.37393
```

```
Linear Regression

151 samples
  6 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 150, 150, 150, 150, 150, 150, ...
Resampling results:

  RMSE      Rsquared   MAE
  89.37393  0.5890984  72.52832

Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
#summary(model3_test)
```

På baggrund af vores MSE værdier, har vi som sammenlignings grundlag opstillet et histogram der visuelt viser, hvordan MSE værdierne bliver mindre i takt med at modelkompleksiteten stiger.

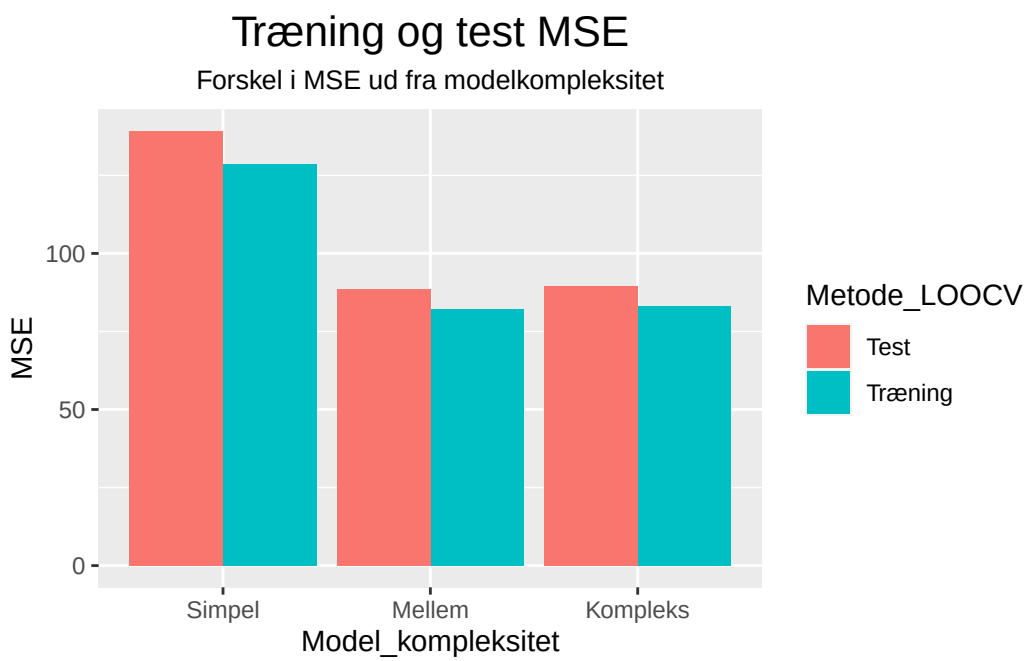
```
Metode_LOOCV = c("Træning", "Træning", "Træning",
                 "Test", "Test", "Test")
Model_kompleksitet = c("Simpel", "Mellem", "Kompleks",
                       "Simpel", "Mellem", "Kompleks")
MSE = c(128.7495, 82.09959, 83.20509, 139.1997, 88.55084, 89.37393)
```

```r
test_frame = data.frame(Metode_LOOCV, Model_kompleksitet, MSE,
                                stringsAsFactors = TRUE)
test_frame$Model_kompleksitet <- factor(test_frame$Model_kompleksitet,
                        levels = c("Simpel", "Mellem", "Kompleks"))

#test_frame



MSE_plot <- ggplot(test_frame) +
 geom_bar(aes(x = Model_kompleksitet, y = MSE, fill=Metode_LOOCV),
 stat = "identity", # Ikke transformere data.
 position = "dodge") +
 labs(title = "Træning og test MSE",
 subtitle = "Forskel i MSE ud fra modelkompleksitet") +
 ggeasy::easy_center_title() +
 theme( plot.title = element_text(hjust = 0.5, size = 16),
 plot.subtitle = element_text(hjust = 0.5, size = 10),
 plot.caption = element_text(hjust = 1, face = "italic", size = 10 ))
MSE_plot
```

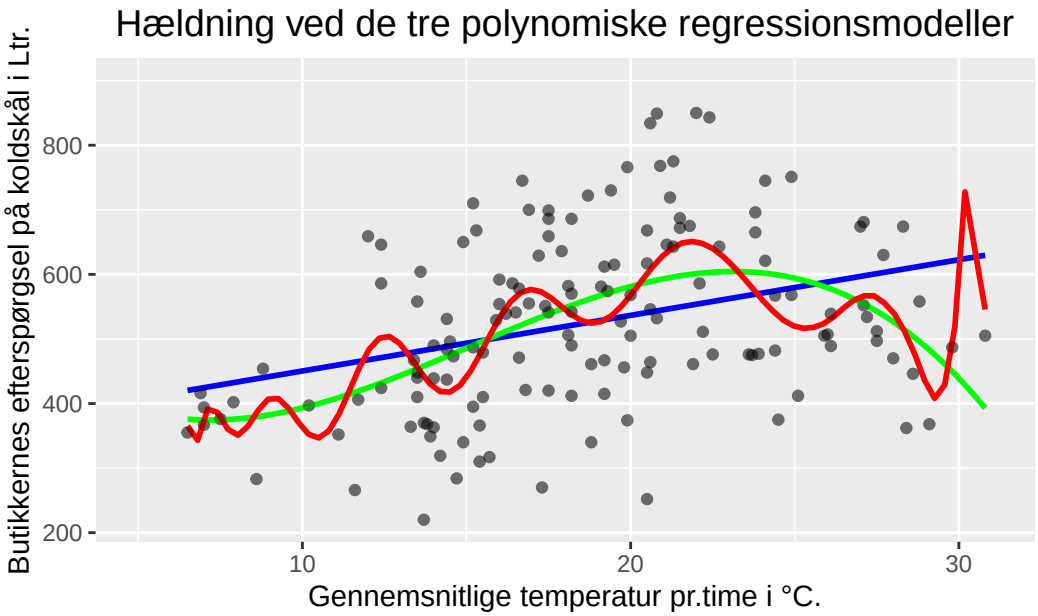



```
#test_frame
#str(test_frame)
#levels(test_frame$Kompleksitet)
#attributes(test_frame)
#glimpse(test_frame)
```

```
x <- data3$temp_mean_past1h
y <- data3$efterspørgsel
data <- data.frame(y, x)

ggplot(data3, mapping = aes(x=x, y=y)) +
  geom_point(alpha=1/3) +
  geom_smooth(method="glm", formula = y ~ poly(x, 1,
                                               raw=TRUE),
```

```
                                                    se=FALSE,
                                                    colour="blue") +
  geom_smooth(method="glm", formula = y ~ poly(x, 3,
                                                    raw=TRUE),
                                                    se=FALSE,
                                                    colour="green") +
  geom_smooth(method="glm", formula = y ~ poly(x, 22,
                                                    raw=TRUE),
                                                    se=FALSE,
                                                    colour="red") +
  geom_point(data=data3, mapping = aes(x=x, y=y), alpha=1/3) +
  labs(title = "Hældning ved de tre polynomiske regressionsmodeller",
  caption = "Kilde: Tal fra DMI 2002 fra perioden 1/4/22-30/8/22",
y = "Butikkernes efterspørgsel på koldskål i Ltr.",
x = "Gennemsnitlige temperatur pr.time i °C.") +
  ggeasy::easy_center_title() + # Centrerer titlen.
  theme( plot.title = element_text(hjust = 0.5, size = 14),
  plot.subtitle = element_text(hjust = 0.5, size = 14),
  plot.caption = element_text(hjust = 1, face = "italic", size = 10 )) +
  xlim(5, 31) + ylim(220, 900)
```



Kilde: Tal fra DMI 2002 fra perioden 1/4/22−30/8/22

5 Udrulning af anbefalingerne

For at komme med kvalificerede anbefalinger bruges der et Data Product Canvas til, at kortlægge de vigtigste anbefalinger som Thise skal implementere for, at forøge deres datamodenhed. Vi tilbyder fem overordnede anbefalinger:

1. Opgrader deres data science stack til en mere moderne version. Start med at opgradere Navision til også at være cloud-baseret istedet for udelukkende at bruge det via. Windows styresystemet.
2. Opstil nye regressionsmodeller ud fra alt deres historiske produktions data. Brug den multible lineære regression fra analysen som modelskabelon og reproducer den i en anden produktionskontekst.
3. Få medarbejderne til at opfatte sig selv som værende en del af en moderne datamodenhedskultur. Første skridt er at ansatte en dataanalytiker som skal

accelerer datamodenhedsprocessen fra fase til fase to. I fase to begynder man at strukturere dataopsamlingen for at lukke risikohuller der er når forskellige systemer skal indgå i en synergi.

4. Få ledelsen til at være spydspidsen når når datamodenhedsprocessen skal øges fremadrettet.
5. Brug R-Studio som programmeringssprog fordi det er gratis og det arbejder godt sammen med andre programmer som fx. SQL, hvilket Thise allerede som en del af deres data science stack.

6 Konklusion

6.1 Sessioninformation

For at højne gennemsigtigheden printes der en udskrift om den nuværende R session:

```
#sessionInfo(package = NULL) # Udskriver en liste om denne R session.
```

7 Litteratur

Bækby, R. og Kølsen, C. (marts. 2017). *„Find din vej i dataindsatsen"*. I: Alexandrainstituttet.

Hastie, T. og James, G. (august. 2021). *„An introduction to statistical learning"*. I: Springer. 2 udgave.

Holland, S. (jul. 2022). *„Vejret afgør sommerens mængde af koldskål"*. I: Fødevareforbundet.

Jensen, M. L (jul. 2022). *„Thise skruer gevaldigt op for koldskålsproduktionen"*. I: Tv Midtvest.

Kjer, U. (sep. 2022). *„Sommervejret var 3 pct. bedre end sidste år"*. I: Mejeriforeningen.

Osterwalder, A. og Pigneur, Y. (2010). *„Business Model Generation "*. 1. udgave. John Wiley & Sons.

Picopublish (feb. 2022). *„Datamodenhed handler om at blive bedre til at anvende egne data i værdiskabende sammenhænge"*. I: Picopublish.

8 Bilag

Bilag 1 - Interviewguide.

Vi introducerer os	"Vores navn og studieretning"
Formålet med undersøgelsen	"Formålet med undersøgelsen er at undersøge, hvordan Thises produktionsafdeling kan blive bedre til, at forudse produktionen af koldskål ved hjælp af at indsamle den relevante data og derefter omsætte den til prædiktion. På denne måde vil Thise nemmere kunne planlægge deres produktion og derved effektivere og minimere spild af råmateriale og tid."

Rammerne for interviewet

Tidsramme	"Vi forventer, at interviewet tager ca. 30 minutter."

Oplys om GDPR	"Må vi optage samtalen på en telefon? Optagelsen bruges som hjælp til, at huske hvad der er blevet sagt og vil indgå I vores projektarbejde. Lydfilen slettes efter vores opgave er blevet bedømt i januar."

Anonymisering	"Interviewet vil blive behandlet fortroligt og gemmes til efter vores opgave er blevet forsvaret og bedømt og herefter vil materialet blive destrueret. I vores opgave vil jeres udsagn blive anonymiseret, såfremt I ønsker dette."

Ønskes anonymisering?	Ja	Nej

Rollefordeling	Interviewer: Kenneth. Notar: Eva er notar. Ordstyrer: Sanne er ordstyren.

Redegørelse	I forbindelse med vores semesterprojekt, ønsker vi at belyse Thises nuværende status for opsamling af deres data for, at få klarlagt, hvor lang I processen de er og hvor deres næste indsatsområde bør være for at de kan nå I mål med, at blive mere datadrevne i fremtiden. Det er vigtigt, at få tydeliggjort om den primære udvikling ligger hos Thise Mejeri eller om den ligger hos medarbejderne - i begge tilfælde er det vigtigt, at få klarlagt hvor man er i processen og hvad næste step derfor bør være. I forbindelse med dette interview, skal vi gøre opmærksom på, at i deltager frivilligt og at i derfor altid kan trække jeres samtykke tilbage.

Præsentation	
Vi introducerer os	Navn og studieretning
Formålet med undersøgelsen	"Formålet med undersøgelsen er at undersøge, hvordan Thises produktionsafdeling kan blive bedre til, at bruge deres egne data til at øge indtjeningen"
Oplys om GDPR	"Må vi optage samtalen på diktafon? Optagelsen bruges som hjælp til, at huske hvad der er blevet sagt. Lydfilen slettes efter vores opgave er blevet bedømt"
Gøre informanten opmærksom på brug af diktafon	"Lydfilen fra interviewet vil blive gemt i 1 år, hvorefter det bliver slettet. Dine udsagn vil blive anonymiseret, så samtalen ikke kan spores tilbage til dig. Du kan selvfølgelig til hver en tid afbryde interviewet.". "Interviewet tager ca. 30 min."

Tid **Vores individuelle roller under interviewet**	Sanne vil notere noter ned. Eva holder øje med om Kenneth stiller de rigtige spørgsmål..
Indledning	
Smalltalk **Baggrundsspørgsmål** **Alder** **Uddannelse** **Stillingsbetegnelse** **Tid i ansættelse**	"Small talk for at skabe tillid.." *"Inden vi går i gang med interviewet spørger vi lidt ind til din baggrund"* *"Hvor gammel er du?"* *"Hvilken uddannelse har du?"* *"Hvilken stilling er du ansat i"* *"I hvor lang tid har du været ansat hos Thise?"*

Temaer	
Definition af data.	"Data om koldskål er informationer i form af tal , personlig erfaring, billeder eller tekst"
"Hvordan kan Thises produktionsafdeling forbedre udnyttelsen af data i forbindelse med deres koldskålsproduktion?"	"Hvilken data indsamler i, når i producerer koldskål?"
"Hvordan ser en typisk arbejdsdag ud i produktionsafdelingen, når der skal produceres koldskål?"	"Med data mener vi alt lige fra vægt, temperatur, mængde (kg), tid o.l."
"Hvordan dokumenterer produktionsafdelingen dagens produktion?"	
"Hvad påvirker produktionsmængden af koldskål?"	*"Kan du kort beskrive, hvilke trin medarbejderen laver når vedkommende skal lave koldskål"* • *Maskinbetjening*
"Hvordan ved produktionsafdelingen hvor mange kg. koldskål de skal producere i en given periode"	"Hvordan dokumenterer medarbejderen dagens produktion af koldskål?"

"Hvordan kvalitetssikrer Thise for uforudsete stigninger eller fald i efterspørgslen af koldskål?"	"Hvilke faktorer påvirker produktionsmængden af koldskål?" • Vejret • Kultur • Tidligere produktion "Hvordan registrerer produktionsafdelingen, mængden af koldskål der skal produceres? • Mavefornemmelse. • Tal. " Hvordan sikre i kvaliteten af koldskålen i løbet af året?" • Ændre smag/konsistens sig jf sommer/vinter? "Har koldskålen samme smagsprofil hele året?" "Hvilken type mælk anvender i til produktionen af koldskål?" "Hvad gør Thise, hvis ikke de kan få leveret de "produkter" de skal anvende i forbindelse med stor efterspørgelse af koldskål?" "Hvor lang tid tager det, at producere koldskål fra råmælk? (Tykmælk/kærnemælk)
Tema: Datamodenhed.	- *" Hvordan oplever i, at medarbejderne reagere, når de skal ændre deres processer?"* *- Analog vs digital arbejdsgang* *"Oplever i, at medarbejderne er åbne overfor, at skulle tracke deres processer i et it system"* *"Hvordan vil Thise sikre, at medarbejderne får udfyldt den data der er behov for, for at blive mere*

	"Hvordan vil medarbejderne have det med, at de får besked fra en computere ang. dagens produktion?"
Værdier	*"Hvordan modtager medarbejderne informationer om generelle ændringer i deres arbejdsprocesser?"* *"Oplever du, at medarbejderne ønsker, og blive mere datadrevne?"* *"Hvad gør Thise for, at gøre deres medarbejder klar til, at være mere datadrevne?"* *"Har medarbejderne en forståelse for, at ved, at tracke nogle forskellige målepunkter, kan det gøre deres hverdag nemmere?"*
Ressourcer og aktiviteter	
Social virkelighed	
Afrunding	
Opsummering	
Gentag anonymitet og Gdpr.	

Bilag 2

Interview af Søren Jensen fra produktionsafdelingen i Thise Mejeri

Gruppe 1

3. november 2022

Søren fortæller at de producerer helst over midnat så de kan skrive den nye dato på pakken. Han fortæller at de producerer ud fra en statistik og hvis de skal forudsige produktionen af koldskål, så skal de være en form for metrologer. De skal være ekstra opmærksomme på statistik fra tidligere år da klimaforandring gør at

sommeren har forandret sig. Salget er ofte højest i starten af sommerperioden og Thise indsamler data på dette.

Hver dag omkring middag kommer ordrerne fra Coop og produktionen håber at det stemmer nogenlunde overens med det de har tappet. Hvis ikke skal de lav en ekstra tapning ellers kan Coop godt godkende det til næste dag hvis det er en lille procent. Der skal minimum være 8 terminaldage når Coop får det leveret.

Koldskål er utrolig afhængig af vejret. Udefrakommende faktorer har også en afgørende rolle; eks. Hvis Arla sælger koldskål til 5 kr. Thise har mange loyale kunder og et godt image men store besparelser fra konkurrenter kan være afgørende for salget hos Thise. Når Thise har et nyt produkt de vil afprøve, så starter de i Irma og hvis det lykkedes, så går det videre til Coop. Thise sælger kun helårs koldskål i Irma og i uge 27-36 sælges det andre steder.

Koldskålproduktionen foregår ved at man tager kærnemælk over og syrner det, så ryger det i en tank som snurrer rundt og der hældes fløde i og kærnemælken tappes af og blandes med en frugtmix af vanilje og æg. Ca. 15% mix og 85% kærnemælk. Der kan produceres ca. 5-6 ton koldskål i timen på 1 tap.

Kærnemælkdrengen (en erfaren mejerist) sørger for at ph-værdigen er lav nok og produktionen smager også på produktet. Laboratoriet udvælger den første og den sidste og 3 i midten til at smage på hver dag og giver karakter på en 15-trins skala. På den måde kan de nå at stoppe salget hvis det smager dårligt. Karakteren gives for ydre, konsistens, lugt og smag. 15 er fantastisk.

Ifølge Søren, så er medarbejdernes it-kompetencer i produktionen ikke høj.

Medarbejderne i Thise har typisk været ansat i firmaet i lang tid. Søren har arbejdet hos Thise siden konfirmandsalderen.

Bilag 3

- Interview med salgschef Peter Pedersen hos Thise Mejeri

  Gruppe 4

  Interview information:

  Informanten:

  Peter Pedersen: salgchef Asien

  SP 1:

  Hvad underbygger i jeres markedsføring i Asien på?

  Svar:

  Der laves overhovedet ingen markedsføring i Asien. Det er B2B – dvs ingen kontakt med slutkunder.

  SP2:

  Hvis I fik data, hvordan vil I markedsføre jer til det Kinesiske marked for eksempel?

  Svar:

  Det Kinesiske marked er meget forskelligt fra det Danske marked. Meget af marketing er trukket over på de sociale medier. Det er meget, tungere på "influencer" end i den vestlige verden. Kinesiske influencers tjener mange

flere penge end vestlige influencers. Det er et helt andet marked man skal penetrere end, hvis man gerne vil til Tyskland. I Tyskland kan man reklamere i fagblade, f.eks. i økofagblade.

Hvis man skal slå sig igennem Asien så er det en rigtig god idé at få de her influencer igennem som er på forskellige platforme (f.eks. TikTok) og andre kinesisk sociale medie platforme som vi ikke kender til i Danmark. Det kræver også noget som en organisation at sætte sig ind i hvordan man egentlig får mest muligt ud af de platforme.

Thise har været i gang med det på et tidspunkt at oprette en online "store" på noget der hedder "T-mall" (?). Det kan sammenlignes med Amazon. De kan oprette en Thise butik på T-mall hvor de så kunne sælge oste igennem. Det kom de aldrig videre pga logistiske problemer ift distribution i Kina, fordi Kina er ret stor. Der er mange byer hvor der bor mindst 10 millioner mennesker. Hvor skal man distribuere ud fra hvis, man gerne vil ud til Tangchon (en by i Kina? Jeg ved ikke hvordan det staves) som, man byggede Jeg? Har været i med tolv millioner mennesker man, aldrig har hørt om før Er. Det nok og have distribution i Beijing hvor, der er to hundrede kilometer fra til Tanchan? Eller skal man også have en distribution i Tanchan. For en lille organisation som Thise er det meget svært. Den idé var derfor droppet det igen. Ellers skal man kunne have cold-chain distribution af mælk, smør og ost fra Thise til Kina – noget der har store udfordringer. Det er svært og meget dyrt.

De har ikke et markedsføringsbudget. De laver produkter med gode historier. De almindelige media (f.eks. Folkeblad) tager historien op, og så laver de

faktisk deres markedsførings for dem. Det er den måde de har drevet Thise Mejeri helt fra starten af. Det er, at de har lavet gode produkter med gode historier som, de ikke har puttet penge i at få det fortalt.

SP3:

Samles der information om hvordan reklamer/historier i Folkeblad har påvirket jeres salg af produkterne?

Svar:

Ikke helt ned på produktniveau. Der er nogen i kommunikationsafdelingen som holder styr på hvor meget omtale de får - om den er positiv eller negativ.

De havde en tilbagetrækning for et halvt års tid siden, fordi der var solgt nogle liter mælk med rengøringsvæske i. Der har ikke været styr på rengøringsprocessen. Der kunne de se at de historier der bliver skrevet rundt omkring var knap så positive som de plejer at være. De "tracer" alt efter om det er godt eller skidt omtale, og har som konsekvens af det fjernet soja og måler heraf aktiviteten i omtalen for ligeledes at kunne vurdere om der er stingende positiv eller negativ omtale. Tracer også trafik på sociale medier

SP4:

Hvordan deler I med coop denne salgsinformation, som resultat af jeres samarbejde?

Svar:

Det er jeg faktisk ikke klar over. Vi har en ret unik måde at samarbejde med dem på. Det er egentlig os, der sender varer til Coop før de sender en

ordre til os. Vi forecaster på, hvor meget Coop skal have, og så sender vi det derover. Der bliver sendt varer tilsvarende forecasten om morgenen, hvor Coop så sender deres ordre i løbet af eftermiddagen. Estimeret 90% af de varer Coop bestiller er allerede sendt, da Thise forecaster korrekt. Derfor bliver det de supplerende/manglende mængder, der leveres om aftenen.
Forud for dette er der en række data, der er tilgængelig for Thiese fra Coop, blandt andet lagerbeholdning. Dog er informanten ikke klar over de konkrete metoder til at samle og modtage disse data.

SP5:

Hvor opsamler I jeres salgsdata fra Coop?

Svar:

Vi har en afdeling, hvor de ansatte sidder med deres forecast – med ca. 6-8 mand, der planlægger produktionen. På grund af den store mængde forskellige produkter, og derfor varenumre, så er det nødvendigt med en stor mængde data, hvilke er tilgængelige i føromtalte afdeling. Udover Coop (som er ca halvdelen af forretningen), så er der endnu flere datakilder.

SP6:

Er Coop med til produktudvikling? Eller er der andre af jeres kunder, der har indvirkning på udvikling?

Svar:

Ja, de er en aktiv medspiller. Nogle gange har Coop en idé om et produkt de gerne vil have, det kan feks være et eksisterende produkt som de ønsker i en

økologisk udgave, så kan de kontakte os for at høre om det er noget vi kan producere. De er en stærk medspiller, og de fleste produkter bliver født her i Thiese, og så banker vi på ved Coop for at høre om de er interesserede i at sælge dem – hvilket de oftest gerne vil. I mange år har vi haft et meget tæt samarbejde med Irma, som vi har brugt en del til at teste markedet af, fordi Irma butikkerne lever af at have et nyt og spændende sortiment – med stor udskiftning – så når vi har et nyt produkt, så ringer vi mere eller mindre over til Coop og siger, at vi sender det med i lastbilerne og om de ikke vil sætte det på hylderne i Irma. Hvis det sælger der, så udvider vi til Superbrugsen og Kvickly på Sjælland, og hvis det så også sælger der, så fortsætter det. De er mere åbne for nye produkter på Sjælland. Det har været fremgangsmetoden i 30 år, men processen er mere strømlinet pga datakrav etc.

SP7:

Hvordan arbejder I med sæsonvare vs. Ikke-sæsonvare i forhold til salg og marketing?

Svar:

Udbuddet af mælk er sæsonbetonet, så produktionen heraf vil variere. Vi prøver derfor at få solgt alt vores økologisk mælk, så vi rammer en balance, hvor den mælk vi får ind, den bliver også solgt.
En af vores mest sæsonbetonede varer er koldskål, som vi måske producerer højest 6 måneder om året. Vi gør ikke ret meget i marketing – eller jo – der

er nok flere billeder af koldskål i tilbudsbladene henover sommeren end der er nu her (læs: november). Vi har også nogle oste, der er sæsonbetonet, feks: en juleost. Der får ostehandlerne noget markedsføringsmateriale med fra os, når de køber disse oste ind.

SP8:

De salgspunkter, datamæssigt, I får ind fra feks Irma når I tester jeres produkter igennem dem, er det nogle I får direkte fra Coop?

Svar:

Vi får at vide, hvor meget vi estimerede de ville sælge og hvor meget de faktisk sælger. Hvis vi har overestimeret, så får vi faktisk produkterne tilbage igen, og derfor er det meget vigtigt, at vi estimerer korrekt. Det bliver lidt tricky, da vi sender produkter ud inden, at de bestiller noget, hvorfor det er meget vigtigt med en god planlægningsafdeling. Og den afdeling er meget afhængig af data, hvilket de indhenter fra Coop. Koldskål kan være ekstra udfordrende, da selv sådan noget som vejret kan have stor indflydelse på salget. Så det er også noget data vi bliver nødt til at tage i mente.
Vi laver også ismix – som blandt andet benyttes i Paradis is – som jo også er en høj sæsonvare, hvor holdbarheden er begrænset til 3 uger, hvorfor vejret også har stor indflydelse i forhold til, hvor meget der bliver brugt og derfor solgt.

SP9:

Hvordan bruger I informationsdata til den bedste markedsføringsstrategi?

Svar:

Vi laver ikke meget markedsføring, altså vi bruger penge på bannere og annoncer, men det er en in-house designer, der er ansat til at lave alt markedsføringsmateriale. Men vi køber ikke annoncer på sociale medier, aviser, TV etc. Vi bruger ikke de gængse markedsføringskanaler, men vi har stadigvæk noget markedsføringsmateriale til at kunne fortælle de gode historier. Det er allerede tænkt ind i produkterne, altså historierne kommer faktisk næsten før produkterne og hvis de ikke kommer først, så kommer de sammen med produkterne. Et eksempel herpå er vores fuldmåne ost – historien bag er, at osten udelukkende er lavet af mælk, der bliver malket, når der er fuldmåne. Og det er en rigtig god historie, men osten er blevet så god, at vi ikke kan producere nok. Der kom historien før produktet feks.

SP10:

Hvad er det I helt præcist eksporterer til Asien?

Svar:

Det er 98% pulver. Det er mælkepulver – i forskellige variationer – og valle-pulver. Lidt ost og lidt langtidsholdbar mælk, og disse to ting går til den samme kunde, som bruger dem i en forretning, hvor de sammensætter og sælger gavekurve. Hertil står Thiese også for noget "markedsføringsmateriale" igennem blandt andet flotte billeder. Thiese understøtter mere kunden end at skabe markedsføringen.

Feks i samarbejde med Coop, så er det en del af samarbejdsaftalen, at den anden part indbetaler x antal kroner til at bidrage til markedsføringen af ens produkter.

SP11:

Har du nogle idéer til, hvordan Thiese kan gøre sig mere datadrevne?

Svar:

Informanten ved, hvor godt data det er og at data benyttes til alt. I selve produktionen alene, så har man en masse data man er afhængig af, da det er råvare, der skal omsættes til færdige produkter. Fuld traceability skal altid være være mulige for os at lave her i Thiese.
Informanten er i tvivl om, hvorledes der er datapunkter, hvor man kan optimere arbejdsgangen eller processer.

SP12:

Hvad er jeres vision for jeres firma? Hvor vil I gerne være om feks 5 år?

Svar:

Thiese har en ejerform, hvor det er landmændene, der ejer virksomheden. Så visionen er i højere grad at fortsætte eksistensen fremadrettet med de grundværdier der er i Thiese. Vi vil gerne gøre en forskel. Vi er ejet af vores andelshavere, som også er dem, der leverer mælken og råvarene. Thiese har altid været en ordentlig virksomhed, som behandler naturen, kunder og ansatte ordentlig. Visionen er at lave gode, økologiske produkter, og blive

bedre til at passe på planeten, fordi mejeridrift ikke er den bedste måde at producere fødevare på. Muligvis kigge mere på plantebaserede produkter, fordi vi er langt fra at være CO2 neutrale i dag, da vi har så mange køer. Her er der faktisk et område, hvor vi kunne bruge hjælp igennem data (læs: CO2), fordi det er noget kompleks.

En ting er, at vi skal have vores traceability ind i en blockchain, da det er godt for slutbrugeren – uanset deres geografiske placering – så det er åbent for alle at se, hvordan deres produkt er blevet leveret til deres lands-/verdensdel. Der kan man også lagre CO2 i den blockchain – altså hvor stort dens aftryk er – der tænker informanten godt, at man kan bruge data til at skabe større gennemsigtighed.

Interviewer:

Det lyder til at I har indblik i, hvor meget CO2 der produceres ved jeres landmænd og indtil det kommer til jer

Informant:

Vi ved godt, hvor meget brændstof vores leveringsmetoder benytter, men det her med at få det ned fra enkelte transportmetode til enkelte liter mælk, kunne være utrolig interessant. Så er det tilgængeligt for slutbrugeren på en anden måde, og mere troværdigt end, hvis det kommer direkte fra os.

Bilag 4

Interview med Bjarne Justensen Senior Demand planner hos Thise Mejeri

Gruppe 5

Hvad er det salg og marketing funktionen laver her hos Thise?

"Jeg sidder hos demand planning, og ser ikke på salg og marketingdelen. Jeg har ikke nogen rigtig ide om hvad de laver. Jeg sidder med demand planning, og ikke så meget ud af huset opgaver. Jeg modtager fx nogle data fra Coop og ser på de historiske data for at danne mig et indtryk af tingenes tilstand, og hvordan jeg forventer det vil forløbe"

Hvordan bliver jeres data i dag, indsamlet fra jeres kunder?

"Det bliver indsamlet i Navision og består af salgs data, forecast data kommer fra et forecast ark fra Coop bland andet, vores grossist salg er historiske data, og så deres kampagner af de enkelte kampagner. Det er ikke så omfattende kampagnetræk."

Hvilke typer af informationer er det som der samles i Navision?

"Det er altid salg data, når vi laver vores forecast ryger data tilbage i Navision."

I forhold til salg, hvad er så de vigtigste data som i kigger på, og indsamler i Navision når i fx skal forudsige salget?

"Jeg kigger meget på de tidligere ordres historik, og ser på hvordan tendenserne var i året. Vi er meget baseret på hvordan kurverne forløber på året, for at kunne sige hvordan kurverne vil forløbe det igangværende år. Så kigger jeg på hvordan tendensen er for øjeblikket, og så vil jeg følge den tendens der var tidligere med det niveau som vi ligger på for øjeblikket. Det er den metode jeg bruger til at lave forecast på."

Hvem har adgang til de her data i indsamler, er det mest dig som sidder med de her tal og informationer og laver nogle rapporter, eller er det noget som den menige medarbejder har adgang til og bruger?

"Når jeg laver de her rapporter, sætter jeg det ind i Navision og så hiver produktionsplanlæggeren dem ud af Navision for at lave en produktionsplan ud fra det."

Er det alle medarbejdere i produktionen som så har adgang til det?

"Det er kun produktionsplanlæggeren. De vil så arbejde videre med rapporten og herefter giver de det/den videre til næste led i produktionen og så videre."

Hvor godt vil du vurdere (hvis du kan) at Thise mejeri er i denne her datamodenheds process? Fra en skala på 1-5, hvor 1 er vi kun har monitering men ikke noget man altid træffer beslutninger ud fra. Hvor 5, er fuldautomatiske processer samt at hele strategien hos Thise Mejeri ligger til grunde i data.

"Jamen, vi er sådan midt i mellem. Der arbejdes stærk hen imod at blive datadrevne. Det vil sige, at alt den forecast jeg laver ryger tilbage ind i Navision som vi(produktionen) så producerer efter. Så alt historisk data, og kampagne flow bliver lagt ind i forecastet og bearbejdet af mig, for at sikre at det er de rigtige mængder vi producerer til den rigtige uge. Så på den måde er vi ret langt i den del."

Og er det noget som du manuelt skal skrive ind, eller sker det automatisk?

"Nej, vi har en udbyder som sidder og kigger på historiske tendenser, som så kommer med et forslag til hvordan forecastet kan se ud. Så kigger jeg det så igennem, først på en overordnet plan, jeg kigger på nogle grupper fx mælk, for at se om tendensen ser rigtig ud her. Ser det så rigtigt ud, kigger jeg lidt dybere ned i forslaget, fx herefter kigger jeg ind i sødmælk, ser det så rigtig ud kigger jeg ned i fløde. Som udgangspunkt får vi forecast fra en underleverandør, som løbende modtager vores salgs data, og så laver de forecast ud fra historiske data og ligesom jeg nævnte før, så viser det flowet og tendensen fra hvordan det har været for at finde ud af hvor vi skal ligge henne nu."

De salgstal i får fra fx COOP og deres forskellige kæder, er det noget som bliver sendt direkte til jer?

"Ja det bliver sendt direkte til os. Vi får kampagneplaner for en periode 10-12 uger frem."

Du har været lidt inde på hvad jeres ambitioner er med henblik på at blive datadrevne. Hvad vil du mene at der skal til for at det kan ske? Er det mere fra ledelsen det skal komme, eller er det mangel på kompetencer hos medarbejderne?

" Jeg vil mene at det lige så meget handler om kompetencerne. Altså jeg er ikke superuddannet, jeg er jo bare autodidakt, demand planner. Det betyder

jo en del for hvordan man ser på data. Selvfølgelig har man nogle andre kompetencer når man ikke er uddannet, men jeg er ikke uddannet datamatiker eller lignende, der har vi(Thise Mejeri) en lille brist der i forhold til at vi godt kunne bruge nogle som jer (dataanalytikkere)."

Spørgsmål fra Izels gruppe:

Du nævner at I får forecast. Vi har et projekt om koldskål, hvor vi gerne vil lavet et forecast, hvad ville du kigge på at for at lave en forecast hvis du skulle kigge på det?

"Når jeg helt praktisk sidder med det, har jeg 3-4 parameter jeg ligger sammen for kunne forecaste. I dette tilfælde hvis jeg skulle sidde og forecaste 12 uger frem, er det lidt begrænset, da jeg vil mangle data for vejrudsigten, jeg mangler at vide noget om beholdningen for det enkelte terminaler. Så der sidder jeg bare og kigger historiske data, fx hvordan så det ud i uge 20 sidste år, der solgte vi måske 21.000, så ville jeg skulle bruge min mavefornemmelse til at forecaste hvordan tendensen så ville se ud i år. Men går man helt tæt på, fx ugen før, så ville jeg kigge på hvad har vi af beholdning, hvordan er vejrudsigten og den langsigtede vejrudsigelse, bliver det godt vejr? Så hvad har vi her på lageret på Thise. Men overordnet set ville det være historiske data jeg baserer forecastet på, hvad plejer man at gøre i denne uge. Koldskål er et sinddsygt dårligt eksempel, da det er så vejrbestemt, så på trods af at man kørte en kampagne, er det ikke sikkert at man ville ramme tæt på hvad der var forventet hvis vejret er dårlig. Det kan også gå den anden vej, fx man siger man har et budget på 10.000kr på kampagnen, men den sælger for 30.000kr. Koldskål er så svær, da den er så vejrafhængig. Vi kigger også

på konkurrenterne, fx hvad sælger Arla den for, sælger de den for 10kr og vi ville have solgt den for 18kr så vil folk vælge den til 10kr. Især i tider som nu hvor folk har færre penge, vil de altid vælge det som er billigst."

Hvad ville så være et bedre eksempel?

"Et eksempel kunne være græsk yoghurt, den er sæsonpræet, man bruger den om sommeren bland andet til Tzatziki. Ved uge 16/17 vil efterspørgslen/salget begynde at stige indtil uge 32 og begynde at falde igen, så vil det gå lidt ned, og så stiger det lidt igen da der kommer lidt mere fokus omkring til jul. Det er måske sådan en jeg hellere ville have kigget på end koldskål, da den stadig er sæsonpræget men også mere forudsigelig end koldskål, og man er uafhængig af vejret udenfor.

Koldskål er så dårlig, da den er så uforudsigelig, da vi på en dårlige uge måske kun sælger 10.000, men på en god uge op til 70-80.000 stk., hvor det eneste som påvirker salget, er vejret"

Jeg kan forstå i ikke rigtig laver så meget markedsføring, men at det mere er historien som føler produktet.

"Coop er vores markedsføring, her vil vi(vores produkter) ofte blive præsenteret store og der vil tit og ofte stå en kort historie om vores produkter. Det vil så sige, at det er den måde vi lærer folk Thise at kende."

I siger i har information om hvor meget i bruger på en given kampagne i forhold til salget?

"Vi får at vide fra Coop at SuperBrugsen og Kvickly kører samme kampagne avis, og at de vil køre en kampagne i uge 42. Så får vi af vide hvor meget Kvickly forventer at sælge ekstra i den uge ud over deres normal salg, og det samme for SuperBrugsen. Det ligger vi så ind i vores forecast system, Perito, så kan vi se at der er en forventning til et salg, på vores kurve i forecastet, her kigger vi så på om det ser rigtigt ud med hvad der rent faktisk sker. Vi har data helt ned i detaljer, delt helt ud på dagene. Fx kører Coops kampagner torsdag til torsdag, så skal vi så have hentet en stor del af kampagnebudgettet ind inden kampagne avisen fylder, så butikken er fyldt op inden kampagnen starter om torsdagen.

Vores spørgsmål igen

Hvad er konsekvenserne af forecast hvis det ikke rammer plet, og hvor store kan de være?

"Jamen hvis forecastet ikke er rigtigt, så har vi pludselig ikke den vare som kunden efterspørger. Det er så mistet salg, (hmm) det er svært at svare på hvor ofte det så sker. Det sker ind imellem. Det sker nogle gange at de (Coop) kører nogle kampagner hvor vi ikke er blevet ordentligt oplyst. Fx for nyligt havde vi en fragt til Tyskland, hvor Coop ikke havde meldt ud at de havde tilbudt vores smør, så trækker de pludselig en rigtig stor del smør, det var vi ikke forberedte på da vi ikke havde produceret til dette, og smør tager typisk 4-5 dage før man kan ændre noget. Så når de trækker en stor mængde, så har vi en 2-3 dage hvor vi ikke kan levere til de andre Coop butikker. Så hvis vi ikke er forbedret, og Coop ikke har fortalt os at de har/vil køre en

kampagne, så kommer vi til at underlevere, da det tager tid at producere. Mælk er dag til dag, men syrnet produkter er ofte 2-3 dage."

Hvordan monitorerer i salget, fx hos Coop, om salget hos Coop går godt eller om det går skidt?

"Jamen vi kigger på serviceegrad. Salget sidder jeg ikke så meget med, jeg sidder ikke med salgstal men jeg sidder med mængderne. Vi kigger hver dag på servicetallene, for at se hvordan vi har preformet ud til Coop, vi har en målsætning om at levere 98,5% på servicegraden til alle vores kunder."

"Det er snart 1,5-2 år siden vi startede med Perito forecasting system, og vi lærer stadig rigtig meget, og det gør de også. Det er en proces. I starten havde de ikke så mange historiske salg. Når man fx kører en kampagne så er der typisk en top i denne hvor det går bedst, og denne top ville jo så ikke skulle bruges næste år for at lave en forecast, man skal altså have fjernet denne"unaturlige" top som er skabt ud fra kampagnen. Vi er blevet bedre og bedre til at forecaste. "

Hvad synes du selv, at det kunne være rart at have?

"Det sværeste for mig er ikke nu hvor tingene kører lige ud. Det er den nemmeste periode vi er i gang med nu, da salget efter sommerferien og så frem til sommerferien, er meget ligetil og den samme. Det er specialperioderne, sommerferie, helligdagene hvor vi kan se en stor ændring i vores salg."

Hvad er det som gør det svært at forecaste, i har kampagneflow og software, er det fordi de ikke kan forudsige de her speciale tider. Er det jer som ikke selv kan forudsige eller?

"Det som gør det svært ved fx jul, det er at i år så falder den forskudt fra sidste år, så spørgsmålet er hvilken dage hiver Coop en stor mængde ind, gør de det en dag forskudt fra sidste år eller? Oveni, så har markedet i år ændret sig, fx er salget af änglemark steget 17-18%, det gør det svært at forudsige fx salget af mælk. Det er noget nemmere ved fx påske og pinse, for der ved vi hvornår den falder hvert år. Julen er faktisk den sværeste at forudsige, og forecaste. Det skyldes også at det er svært for at forudsige hvad Coop gør, vi får nogle forecast fra Coop om hvornår og hvad de tror de gør, men det er ikke altid helt sådan at det forløber."

Er det rigtigt at Coop er jeres største kunde? Hvem har i ellers af kunder?

"Ja det er rigtigt. Så har vi nogle forskellige grossister, alle de store bland andet Dagrofa mv. Det er et stødt voksende marked, jeg vil gætte på vi voksede 25% fra sidste år i forhold til salget hos grossister. Det er et godt marked, og godt at samarbejde med store grossister for Thises navn."

Bjarne Justensen

Senior demand planner

Bilag 5

Efterspørgsel efter Koldskål

En fiktiv undersøgelse

Eksamen vinter 2022

Coop Danmarks efterspørgsel efter koldskål et sted i København. Nærmere betegnet i nærheden af en vejrstation tæt på Landbohøjskolen. Vi kommer ikke nærmere ind på helt nøjagtig, hvor butikkerne, som efterspørger koldskålen, befinder sig. Det er ikke så vigtigt.

Derimod er det vigtigt for Thise at kunne udregne Coop's efterspørgsel efter koldskålen for at kunne planlægge produktionen. I denne opgave er formålet at lave en model for efterspørgslen i de pågældende butikker på baggrund af nogle forklarende variabler. De potentielle variabler er angivet nedenfor. Det er vigtigt at overveje, hvilke variabler der skal tages med på baggrund af forskellige kriterier.

Vi ser på en periode fra primo april til ultimo august. Lad os antage, at Coop (for de pågældende butikker i området omkring Landbohøjskolen) bestiller koldskål hver dag. Lad os antage, at de leveres fra et Thise-fjernlager i Stor København tæt på butikkerne.

I skal bruge Validation set metoden, LOOCV, eller en k-fold cv metode. Lav mindst tre konkurrerende modeller, og udvælg den med mindst MSE.

I skal bruge Crisp-DM og starte med at forstå forretningen ud fra kriterier i denne opgaveformulering. I skal konvertere forretningsproblemet til et Data Mining problem. Vise at I mestrer programmering, hente data fra eksterne kilder, lave eksplorativ analyse, og foretage de rigtige analyser. Endelig skal I i en konklusion præsentere resultaterne i overensstemmelse med formålet og på en måde, så de er til gavn for de relevante beslutningstagere i Thise.

Alt skal være reproducerbart. Denne opgave skal besvares i RMarkdowm, og

hver kode chunk skal kommenteres, og I skal kunne argumentere for de valg, I træffer undervejs. Følg punkterne i Crisp-DM.

I skal endeligt foretage nogle prædiktioner på baggrund af den bedste model ud fra et datasæt med relevante x-variabler og tilhørende relevante værdier. Konstruér selv dette datasæt.

Hvilke variabler kunne have betydning for regressionen:

- Mere end 25% af butikkerne i det pågældende område er løbet tør for kammerjunkere: "kammerjunkere" (dummy-variabel; findes i det udleverede datasæt).
- Weekend og helligdage (fredag, lørdag og søndag, og helligdage): Dummy; =1 hvis den pågældende dag er en fredag, en lørdag, en søndag eller en helligdag (dansk).

Den kan I selv lave ud fra datovariablen, som I har fået udleveret.

- Forskellige vejr-variabler: Bl.a. temperatur og fugtighed.

Disse kan I finde hos DMI og kan tilgås via en API. Variablerne merges med de andre variabler. Brug en datovariabel som key-variabel.

Potentielle variabler: "temp_min_past1h", "humidity", "temp_dry", "temp_dew","temp_max_past1h", "humidity_past1h" og "temp_mean_past1h". https://confluence.govcloud.dk/pages/viewpage.action?pageId=26476616

Når I anvender disse variabler, skal I overveje om alle variabler er interessante. Giver det mening at droppe en eller flere af variablerne. Hvilke problemer

vil det give at tage alle variabler med? Er der en lineær relation mellem efterspørgslen og variablerne?

Hint: Hent kun observationer genereret klokken 12:00 hver dag.

Hint: Der er måske ikke en lineær relation mellem temperatur og y variablen (efterspørgsel).

- Har temperaturen de sidste tre dage været over 25 grader ved middagstid? Dummy-variabel. Den skal I selv lave på baggrund af data fra DMI.
- Forventet lagerbeholdning i butikkerne: Er en kategorisk variabel, der kan antage værdierne 1=lav, 2=mellem og 3=stor. Denne variabel vil fremgå af det udleverede materiale: forventet_l_lager.
- Endeligt er måned måske også relevant. En sådan variabel kan I også lave ud fra datovariablen.
- Hint: Nogle af den genererede variabler kan måske med fordel konverteres til factors!
-

Bilag 6

Virksomhedscase: Thise

Thise Mejeri

Thise Mejeri har, over de senere år, arbejdet struktureret med at løfte deres IT-platform, og har nu et ønske om at blive mere datadrevne. De har løbende

indsamlet en del forskellige data, og er nu interesseret i værdiskabelse fra den data. Til det formål er der etableret et samarbejde mellem Thise Mejeri og Dania's PBA i dataanalyse om deres 1. semester projekt.

Problemfeltet er derfor; "Hvordan kan Thise Mejeri blive mere data dreven?". Dette skal specificeres yderligere af de enkelte grupper.

Projektet er et gruppeprojekt med givne grupper på 3-4 personer, som følger Dania projekt guidelines (se i moodle for detaljer).

Vigtige Datoer:

- 03/11 besøg hos Thise
- 11/11 Aflevering af problemformulering pr e-mail til CLOL/BJSO/ANRE
- 06/01 Aflevering af projekt i wiseflow
- 11+12/01 Mundtlig eksamen

Til at komme i gang kan følgende emner måske hjælpe:

- Identificere data punkter i Business Model Canvas for Thise Mejeri.
- Identificer branche benchmarks og virksomheder der gør det specielt godt

Faser i projektet:

- Empathize
- Forstå branche, virksomhed og mennesker

- Ideate
- Identificer områder og løsninger der kan understøtte deres vision data drevet
- Prototype
- Lave analyser / løsninger på baggrund af data
- Test
- Storytelling /business model

9 Figurer

Figur 1 - Busines model Canvas.

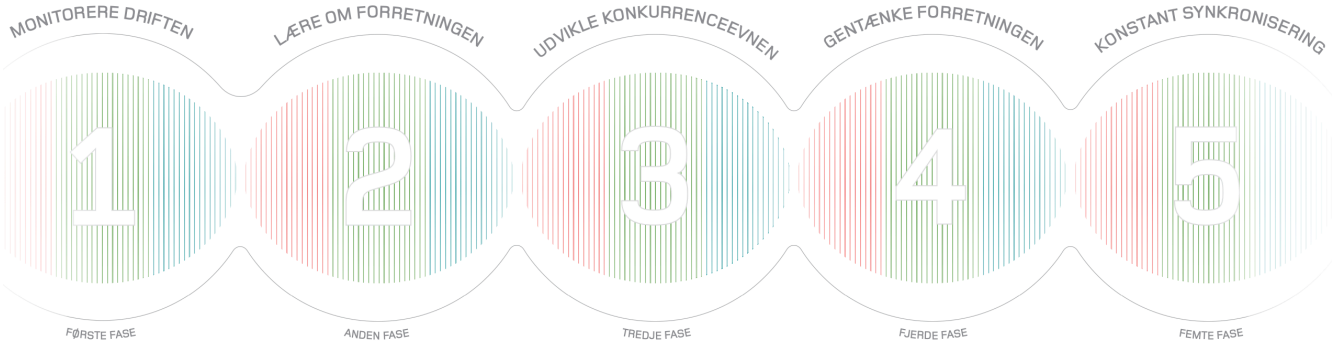
Osterwalder's Business Model Canvas

Thise Mejeri

Key Partners	Key Activities	Value Propositions	Customer Relationship	Customer Segments
• Landmænd (andelshaver) • Kultur og partnerskaber	• Fremstilling af højkvalitets mejeriprodukter	• Thise er et økologisk brand. • Produkter der fortæller den gode historie • Løfte kvaliteten for dyrevelfærd. • Tro imod deres vision • Pris og kvalitet hænger sammen.	• Thises vision for deres produkter og samarbejde med landmændene. • Fortælle den gode historie, som kunderne kan relatere til. • Messebesøg • Kundebesøg	• Coop • Irma • Ostebutik (Thise Mejeri) • Hansen Is
	Key Resourses • Økologisk Jerseymælk • Høj faglighed • Produktionsanlæg • Egne lastbiler		**Channels** • Thise leverer selv deres produkter til COOP • Klausulaftale • Kunden køber varer gennem en VMI(Supply chain management system) terminal. • Pareto til forecasting	

Cost Structure	Revenue Streams
• Produktionsanlæg • Afregning til landmanden • Drift, strøm varme og løn • Råvare • Transport	Thise Mejeri holder hele risikoen forbundet med deres produkt indtjening. Det skal forstås sådan, at når de sender et produkt ud til Coop butikkerne, får de ikke betaling før produktet er solgt. • Øget risiko • Ostebutikken • Coop og Irma • Salgspriserne er fastlagte(Fixed menu pricing)

Figur 2 - Alexandremodellen



Data bruges til at monitorere drift og forretning	Data bruges til at lære om forretningen	Data bruges til at udvikle konkurrenceevnen	Data bruges til at gentænke forretningen	Data bruges til konstant synkronisering

Figur 3 - Data Product Canvas