Prosjektoppgave

Kandidat Nummer: 6 og 26

2022-06-05

Pakker og datasett

Laster inn nødvendige pakker for å få med diverse funksjoner som skal være med i prosjektet.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## Warning: package 'tidyr' was built under R version 4.0.5

## Warning: package 'readr' was built under R version 4.0.5

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(plotly)
```

```
##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout
```

```
library(DataExplorer)
library(knitr)
```

Ved bruk av readr pakken som er i tidyverse pakken kan vi bruke `read.csv()` funksjonen så seperarer datasettet med "," siden csv filer bruker komma for separasjon, for å laste inn datasettene som ble utdelt.

```
AppWichStoreAttributes <- read.csv('AppWichStoreAttributes.csv', sep = ',')
county_crime <- read.csv('county_crime.csv', sep = ',')
county_demographic <- read.csv('county_demographic.csv', sep = ',')
county_employment <- read.csv('county_employment.csv', sep = ',')
WEEKLY_SALES_10STORES <- read.csv('WEEKLY_SALES_10STORES.csv', sep = ',')
weekly_weather <- read.csv('WEEKLY_WEATHER.csv', sep = ',')
```

```
colnames(AppWichStoreAttributes)[4] <- "County_Name"
colnames(AppWichStoreAttributes)[6] <- "Weather_Station"
colnames(WEEKLY_SALES_10STORES)[2] <- "Store_Num"
colnames(weekly_weather)[2] <- "Date"

weekly_weather$Date <- as.Date(weekly_weather$Date, format = "%d/%m/%Y")
WEEKLY_SALES_10STORES %>%
  mutate(Date = as.Date(with(WEEKLY_SALES_10STORES, paste(Year, Month, Day,sep="-")), "%Y-%m-%d")) -> W

WEEKLY_SALES_10STORES$Description = str_to_title(WEEKLY_SALES_10STORES$Description)
```

Oppgave 1: Datasammenslåing

Bakgrunn for oppgave 1

Vi har fått i oppgave å først slå sammen forskjellige datasett. Vi begynte først med å se hvilke kononner som datasettene har tilfelles, og da var det flere som hadde County_Name, Store_Num og Weather_Station tilfelles. Det var enkelte kolonner som hadde like verdier i radene, men ikke samme kolonnetittel, da brukte vi `colname()` for å endre disse slik at alle datasettene med eks 'County_name' verdier hadde samme kolonnetittel. Dette gjøres for å kunne slå sammen disse datasettene til et stort datasett.

Ved hjelp av Tidyverse og Lubridate pakkene kan vi endre noen av datovariablene. Her har vi først brukt `mutate()` funksjonen i begge datasettene der `dmy()` i "Weekly_weather" datasettet og mdy() i "Weekly_Sales" datasettet.

Datasettene er nå klare til å kombineres, og da brukes en `left_join()` funksjon for å få tillagt ny informasjon uten å multiplisere eller duplisere variablene som er like.

```
DF1 <- left_join(AppWichStoreAttributes, weekly_weather, by = "Weather_Station")
DF2 <- left_join(DF1, county_crime, by = "County_Name")
DF3 <- left_join(DF2, county_demographic, by = "County_Name")
DF4 <- left_join(DF3, county_employment, by = "County_Name")
DF5 <- left_join(DF4, WEEKLY_SALES_10STORES, by = c("Store_Num", "Date"))
```

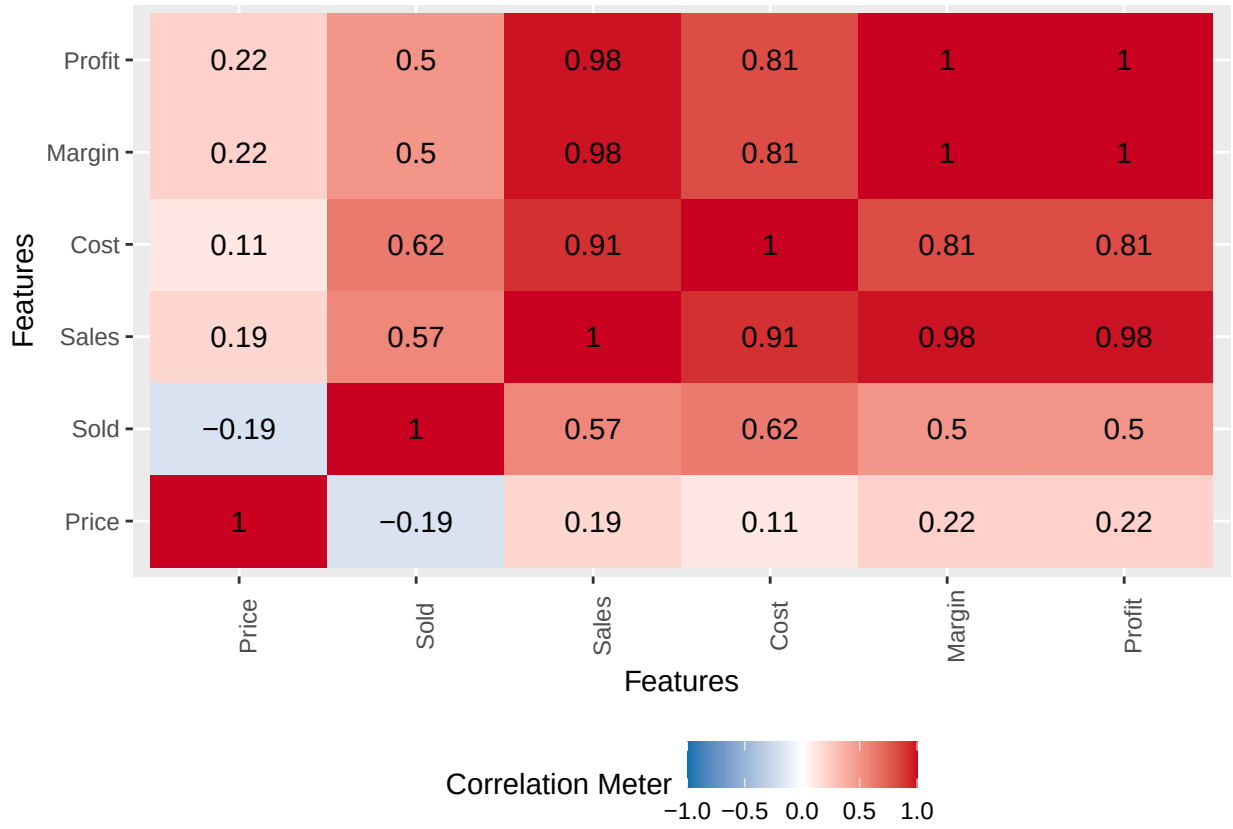
Oppgave 2: Ukentlig salgsrapport

Ved å ha ukentlig rapport vil tillate deg å velge de primære mønstrene i dataene dine. I tillegg vil du også ha tid til å endre strategien din før den faktisk reflekterer på din månedlige rapport

Vi lager et nytt datasett som vi skal kalle for "Power1", her tar vi i bruk funksjoner fra tidyverse til å sortere det nye datasettet. Vi bruker først `Filter()` funksjonen for å velge Uke og Butikknavn, deretter bruker vi `Select()` funksjonen til å velge hvilken variabler som er aktuelle å se på i vår tilfelle er det priser, kosnader, og til slutt bruker vi `Group_by()` funksjonen for å gruppere etter de variablene som vare nr, pris og solgte enheter.

```
Power1 <- DF5 %>%
  filter(Weather_Week == 1, Store_Name == "Power City FreeStand") %>%
  select(Store_Name, Store_Num, Store_City, County_Name, Date, INV_NUMBER,Description, Price, Sold,
         Sales, Tot_Sls, Unit_Cost, Cost,Cost_Percent, Margin, Profit) %>%
  group_by(INV_NUMBER, Description, Price, Sold, Cost, Profit, Margin) %>%
  ungroup()
```

```
plot_correlation(Power1[c(8,9,10, 13, 15, 16)], type = 'continuous')
```



Ved å ta en korrelasjons plot ser vi hvilken variabler som forklarer andre salget og profitten. Med dette har vi da valgt å gå for profitt og solgte enheter.

```
lm(formula = Power1$Profit ~ Power1$Sold)
```

```
##
## Call:
## lm(formula = Power1$Profit ~ Power1$Sold)
##
## Coefficients:
## (Intercept)  Power1$Sold
##     37.5264       0.9126
```

Vi har valgt å bruke en linjær funksjon som har likningen Y = a + bX, hvor verdien av Y er den avhengige variabelen og der verdien av y når x = 0. Modellen sier oss at profitten er den avhengige variabelen (y), som blir forklart av en solgte enheter. Der 1 enhets endring i variabelen x gir b endring i variabelen y. Ved å sette inn en (lm) funksjon som er linær funksjon, får vi korrelasjonen mellom profitten og antall solgte enheter. Der den har en stigningstall på 91,26%.

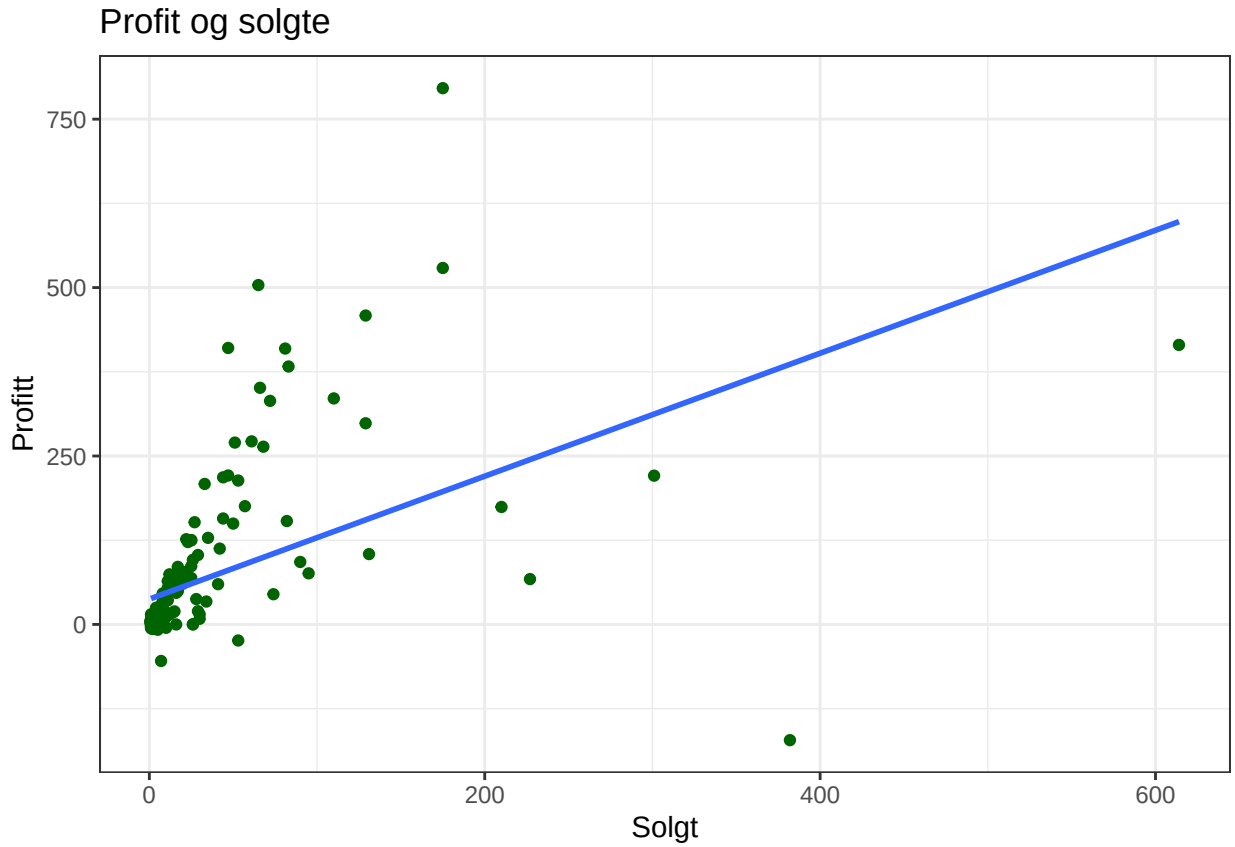
Vi har valgt å se på matvognen "Power City FreeStand" som er lokalisert i fylket Power i Idaho USA. For å få en ukentlig oversikt over butikken og den sine svakheter og styrker, har vi valgt å visuelt vise de solgte enhetene på x-aksen og profitten på y-aksen fra uke 1 i figuren nedenfor. For å få enoversiktlig ukesrapport har vi valgt å sammenlikne profitt mot antall solgte enheter i en uke av de forskjellige varene. Spredningsplotet under viser nettop dette.

Figuren består av solgte enheter på x-aksen og hvor mye profitt de bringer i y-aksen. Den blå linjen er den best egnede linære regresjonslinjen. Det betyr at dette hadde vært linjen som er "to the best fit" for en sum

til punktene rundt. Som vi kan se er et noen verdier som bringer negativ profitt, og dette er gratisprodukter som kan oppstå når butikken har eks "3 for 2, få den billigste gratis" eller "hver 5 sub er gratis"kampanjer. Disse gratisvarene bidrar ikke til å øke profitten til bedriften kortsiktig, men har potensialet til å friste flere kunder, noe som kan være lønnsomt i det langsiktige løp.

```
Power1 %>%
  ggplot(aes(x = Sold, y = Profit)) +
  geom_point(color="dark green") +
  labs(title= "Profit og solgte",
       x="Solgt",
       y="Profitt") +
  geom_smooth(method = lm, se = FALSE) +
  theme_bw()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



I økonomi er prisnivåer en nøkkelindikator og følges nøye av økonomer. De spiller en viktig rolle i forbrukernes kjøpekraft samt salg av varer og tjenester. Det spiller også en viktig rolle i tilbud-etterspørselskjeden. For å få en grundigere forståelse på hvilke prisklasse kundene i Power city benytter seg mest av har vi heltalls rundet alle prisene for så å dele de opp i prisgrupper som begunner på varer til 1 dollar opp til 8 dollar, og utelukke gratisvarene, da de ikke bringer direkte profitt. Dette gjorde vi ved å bruke funksjoner som `if`, `else if` og `else`, har vi laget en funksjon av prisen og katagosert prisklasser.

```
categorize_by_price <- function(price) {
  if (price <= 1.0) {
```

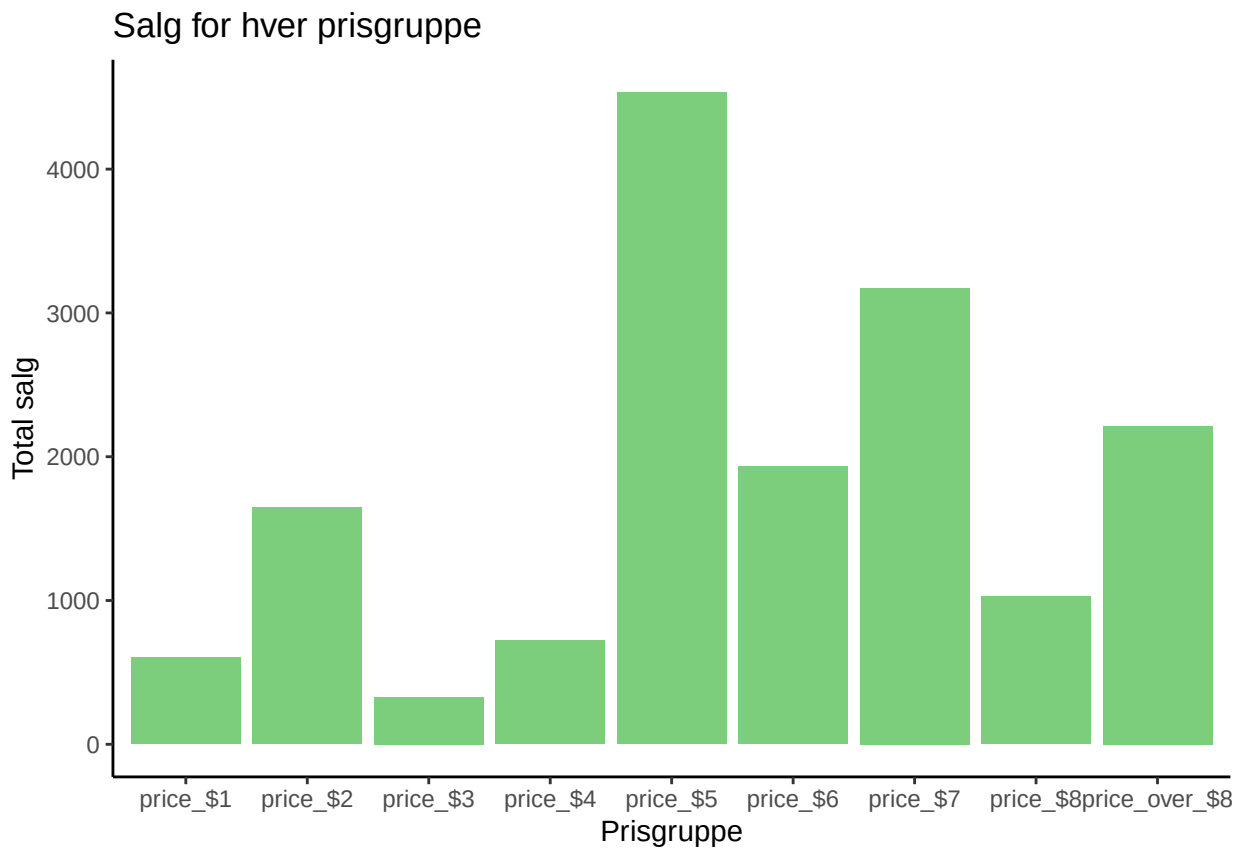
```
    return("price_$1")
  } else if (price <= 8.0) {
    return (paste("price_$", round(price), sep=""))
  } else {
    return("price_over_$8")
  }
}
```

```
Power2 <- Power1 %>%
  mutate(Price_Group = map(Price, categorize_by_price) %>% unlist()) %>%
  group_by(Price_Group) %>%
  summarise(Sold = n(),
            Sales = sum(Sales),
            Profit = sum(Profit),
            Cost = sum(Cost))
Power2
```

```
## # A tibble: 9 x 5
##   Price_Group    Sold Sales Profit   Cost
##   <chr>         <int> <dbl>  <dbl>  <dbl>
## 1 price_$1         32  605.   302.  303.
## 2 price_$2         16 1647.  1044.  603.
## 3 price_$3          9  329.   278.   51.8
## 4 price_$4         18  722.   547.  175.
## 5 price_$5         27 4534.  2973. 1561.
## 6 price_$6         35 1932.  1533.  399.
## 7 price_$7         13 3173.  2210.  963.
## 8 price_$8         14 1029.   726.  303.
## 9 price_over_$8    15 2213.  1746.  468.
```

Ved å bruke ggplot, lager vi en histogram som viser oss hvor mange varer det er i hver prisgruppe. Figur 2 er et visuell presentasjon på hvordan varer til de forskjellige prisklassene selger. Varer som koster ca 5 dollar selger mest. Dette kan være på grunn av mange årsaker, en av disse kunne vært at varene til 5 Dollar er mer fristende for konsumentene.

```
Histogram <-
  Power2 %>%
  ggplot(aes(x = Price_Group, y = Sales))+
  geom_bar(stat= "identity", fill = "palegreen3") +
  labs(title = "Salg for hver prisgruppe", x = "Prisgruppe", y = "Total salg") +
  theme_classic()
Histogram
```



Oppgave 3: Månedlig salgsrapport

En konsernrapport med en mer langsiktig mål. Da har vi valgt å lage to tabeller som viser aggregerte tall fra to forskjellige månder. den ene for april 2012, som er den første månden med data vi har fått, og den andre for januar 2013, som er den siste månden vi fikk utdelt data for. Da har vi en summert tabell som viser antall profitt, solgte enheter, hvor mye det koster å produsere enhetene, og totale salg. Tabeller tar for seg hver av de 10 utsalgsstedene.

```
mndinndeling_april_2012 <- DF5 %>%
  filter(Month == 4) %>%
  group_by(Store_Name) %>%
  dplyr::summarise(across(c(Sold, Profit, Tot_Sls, Unit_Cost),sum))

mndinndeling_april_2012
```

```
## # A tibble: 10 x 5
##    Store_Name              Sold Profit Tot_Sls Unit_Cost
##    <chr>                  <int>  <dbl>   <dbl>     <dbl>
##  1 Lake City StripMall    20466 39130.    5.00      805.
##  2 Littletown StripMall   14947 37848.    5.31      765.
##  3 North Town BigBox      14085 29242.    5.00      934.
##  4 North Town StripMall   18690 42979.    4.85      880.
```

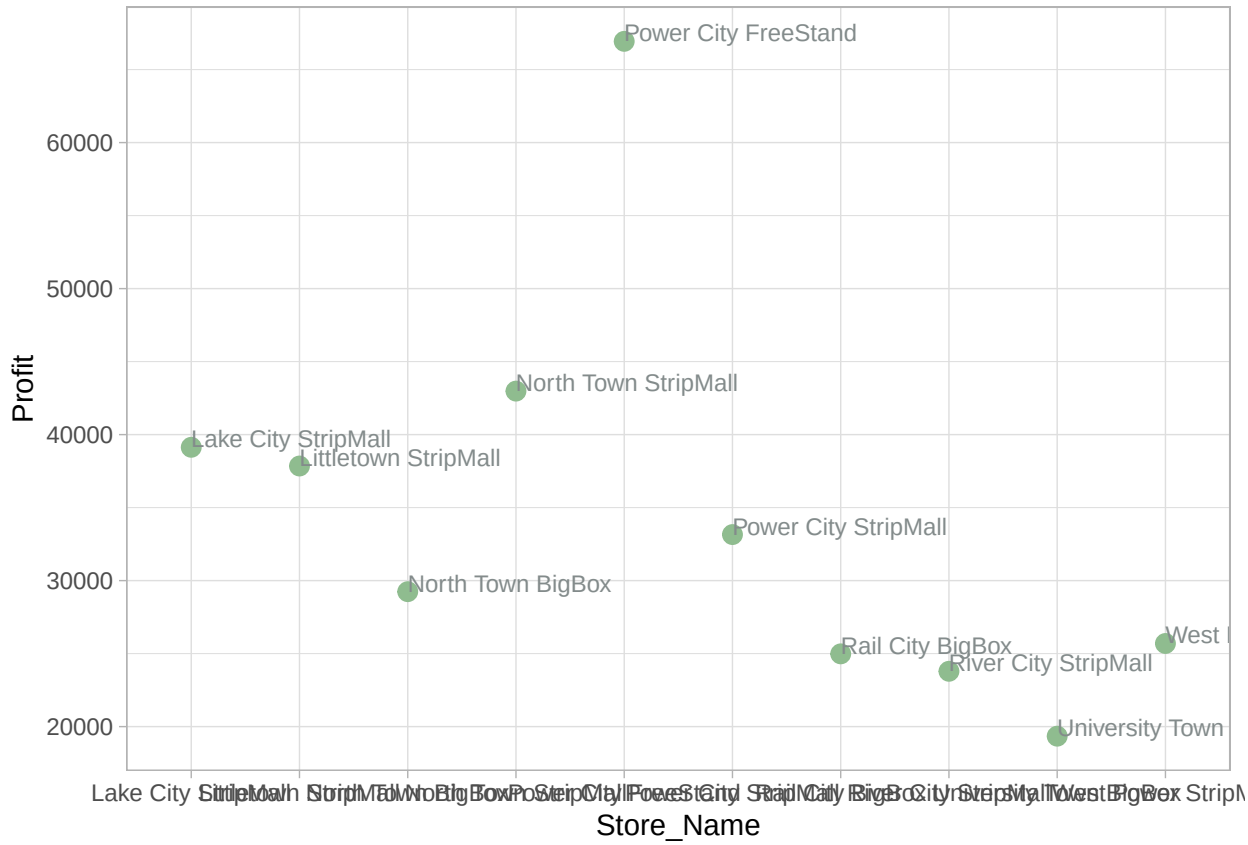
```
##  5 Power City FreeStand   33558 66934.    5.00     1100.
##  6 Power City StripMall   17272 33153.    4.00      726.
##  7 Rail City BigBox       12880 24983.    4.97      798.
##  8 River City StripMall   12431 23789.    5.00      757.
##  9 University Town BigBox  9655 19345.    5.00      724.
## 10 West Power StripMall   12666 25694.    5.00      819.
```

```
mndinndeling_januar_2013 <- DF5 %>%
  filter(Month == 1) %>%
  group_by(Store_Name) %>%
  dplyr::summarise(across(c(Sold, Profit, Tot_Sls, Unit_Cost),sum))

mndinndeling_januar_2013
```

```
## # A tibble: 10 x 5
##    Store_Name              Sold Profit Tot_Sls Unit_Cost
##    <chr>                  <int>  <dbl>   <dbl>     <dbl>
##  1 Lake City StripMall    13208 27698.    4.00      662.
##  2 Littletown StripMall    9819 25081.    4.06      623.
##  3 North Town BigBox       9864 20785.    3.99      613.
##  4 North Town StripMall   14039 31864.    4.00      736.
##  5 Power City FreeStand   19711 43308.    4.00      839.
##  6 Power City StripMall   12232 27577.    4.00      706.
##  7 Rail City BigBox        8314 16640.    3.88      645.
##  8 River City StripMall    6894 15905.    4.00      572.
##  9 University Town BigBox  7846 15963.    4.00      623.
## 10 West Power StripMall    7932 18863.    4.00      576.
```

I spredningsplottet under er alle butikkene i månden april. Dette har vi gjort bevvist for å se hvilken butikk som får mest profitt og hvilken butikk som får minst. x-aksen består av butikknavnene og Y-aksen er den månendlige profitten til hver butikk summert. Her kan vi se at "Power City Freestand" har høyest profitt i månden april og "University town BigBox" får minst profitt.

```
april_profitt <- ggplot(mndinndeling_april_2012, aes(x = Store_Name,
                                                     y = Profit))+
  geom_point(colour = 'darkseagreen', size = 3)+
  theme_light()+
  geom_text(size = 3, aes(label =Store_Name ), hjust = 0, vjust = 0, colour = 'azure4')#+ xlab('Butikk'

april_profitt
```

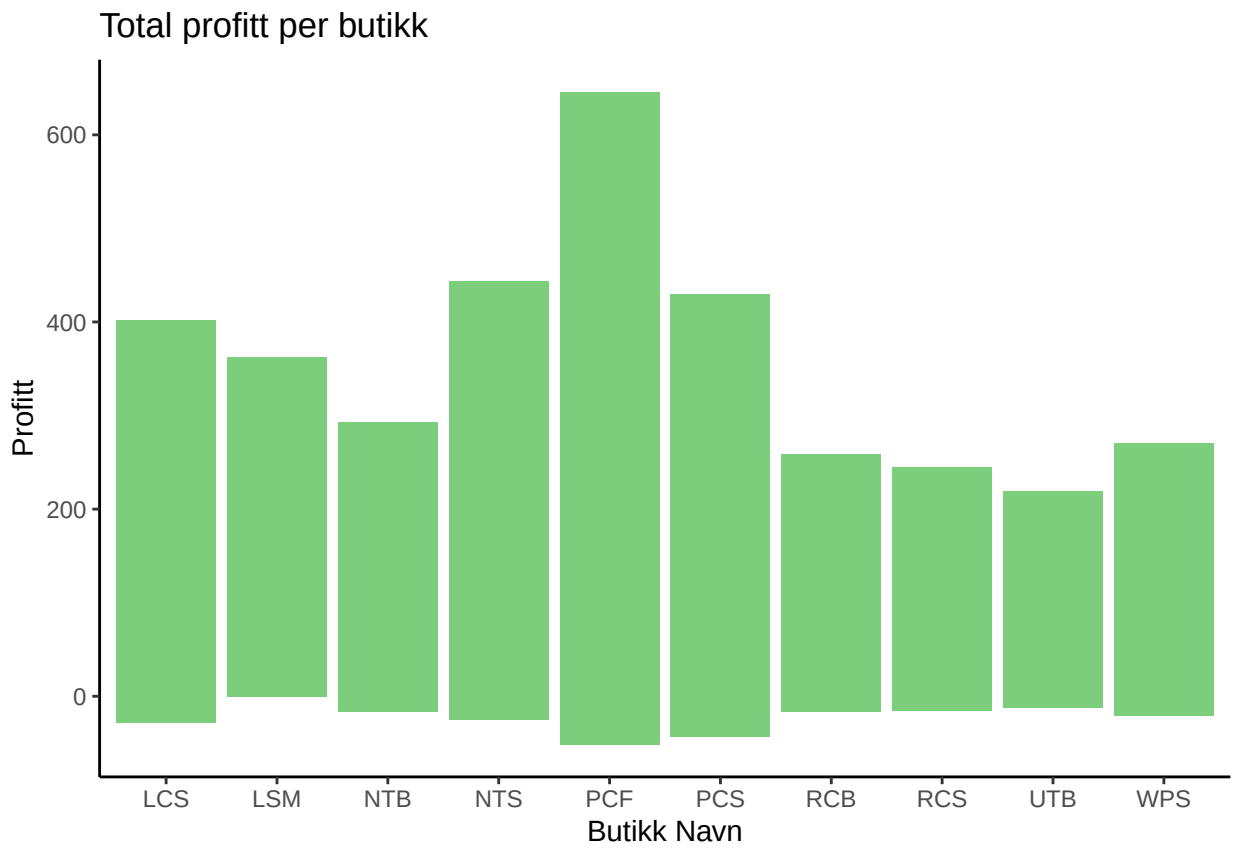


Oppgave 4: Profitt mellom to butikker

I denne oppgaven har vi valgt å se på to forskjellige butikker med forskjellig profitter nivåer, og se på hvilket variabler som spiller en rolle til profitt forskjellene mellom butikkene. Vi velger første butikken som har minst profitt og velger ut variabler som kan være aktuell for å forklare profitt forskjellene i forhold til de andre butikkene. Ved å bruk funksjonen abbreviate så forkorter jeg navnene til butikkene. Ved å lage et histogram ser vi hvilken butikker som har minst og størst profitt. Deretter lager vi et histogram for å vise hvilke butikker som har minst og størst profitt.

```
sss <- abbreviate(names.arg = DF5$Store_Name, minlength = 3, use.classes = TRUE, dot = FALSE, strict = F
                  method = c("both.sides"))

Histo2 <- DF5 %>%
  ggplot(aes(x = sss, y = Profit/1e3))+
  geom_bar(stat= "identity", fill = "palegreen3") +
  labs(title = "Total profitt per butikk", x = "Butikk Navn", y = "Profitt") +
  theme_classic()
Histo2
```

```
## Warning: Removed 2 rows containing missing values (position_stack).
```

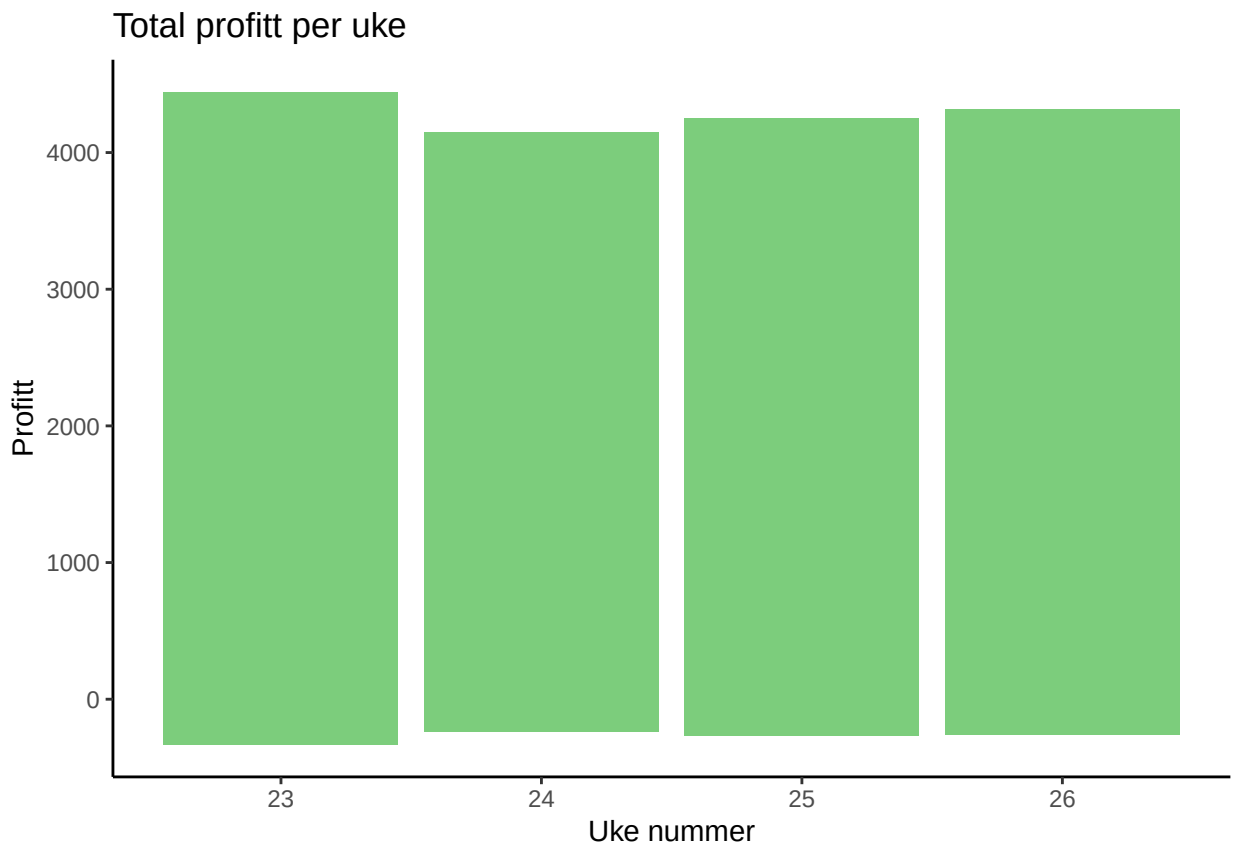


I denne delen tar vi for oss den butikken som har minst profitt og lager et nytt datasett, ved bruk av Filter() funksjonen velger vi butikken "University Town BigBox" , i måneden juni og grupper så etter varene de selger for å få en bedre oversikt. Select() Funksjonen hjelper oss til å velge hvilken variabler vi tar med videre i datasette. Vi lager så et histogram for å vise de fire forskjellige ukene og hvor mye de har i profitt i disse ukene.

```
UTB6 <- DF5 %>%
  filter(Store_Name == "University Town BigBox", Month == "6") %>%
  group_by(Description) %>%
  select(Date, Store_City, County_Name, Price, Sold, Profit, Store_Near_School,County_Crime_Pop ,Store_
```

```
## Adding missing grouping variables: `Description`
```

```
Histo3 <- UTB6 %>%
  ggplot(aes(x = Weather_Week, y = Profit))+
  geom_bar(stat= "identity", fill = "palegreen3") +
  labs(title = "Total profitt per uke", x = "Uke nummer", y = "Profitt") +
  theme_classic()
Histo3
```

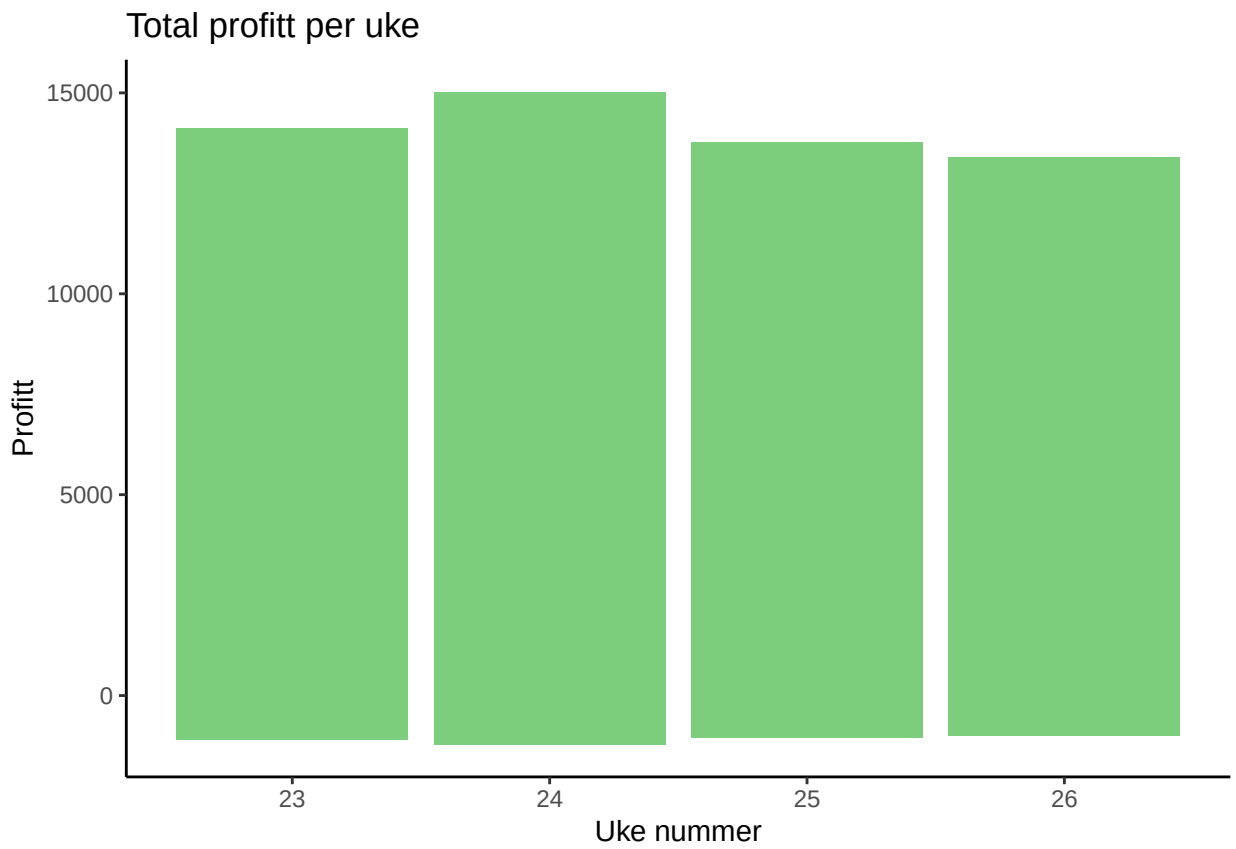


I denne delen tar vi for oss butikken som har mest i profitt av alle 10 butikkene. Vi gjentar samme prosess som i forrige del for å vise samme data i den nye butikken vi har valgt. Ved å bruke `Select()` funksjonen får vi

```
PCF6 <- DF5 %>%
  filter(Store_Name == "Power City FreeStand", Month == "6") %>%
  group_by(Description) %>%
  select(Date, Store_City, County_Name, Price, Sold, Profit, Store_Near_School,County_Crime_Pop ,Store_
```

```
## Adding missing grouping variables: `Description`
```

```
Histo4 <- PCF6 %>%
  ggplot(aes(x = Weather_Week, y = Profit))+
  geom_bar(stat= "identity", fill = "palegreen3") +
  labs(title = "Total profitt per uke", x = "Uke nummer", y = "Profitt") +
  theme_classic()
Histo4
```



Her ser vi at summen av profitten til "BigBox" og "FreeStand" butikkene i de fire ukene som er valgt.

```
sum(UTB6$Profit)
```

```
## [1] 16062.61
```

```
sum(PCF6$Profit)
```

```
## [1] 51944.66
```

Grunnlaget for profittforskjellene kan være både fordi "University Town BigBox" butikken hverken har drive through eller en skole i nærheten av butikken, der "Power City FreeStand" har en drive through og en skole i nærheten. En annen faktor som kan påvirke hvor bra en butikk gjør det i salg kan være antall konkurrenter i nærområdet.Vi ser også at "BigBox" butikken har bare 2 fastfood konkurrenter mens "FreeStand" butikken har 10 Fas-tfood butikker og 25 ikke fastfood i nærområdet, dette tyder på at det er mye konkurranse i nærområdet. Måndesleien er en fast sum som butikkene må betale for å kunne bruke lokalet til salg, og dette er selvfølgelig en faktor som må tas hensyn til dersom man vurderer et nytt utsalgssted. Dersom leien er veldig høy, burde inntektene til butikken være høye nok til å ha rå til leie og andre kosnader samt muligens sitte igjen med profitt. "BigBox" butikken har 15000 USD i årlig leie for lokalet, mens "FreeStand" butikken har 63000 USD i årlig leie for lokalet.

Oppgave 4a: Våre anbefalinger

Ved å se på den minste profittbare og de mest profittbare butikkene ser vi at den mest omsatte butikken har både Drive through og skole i nærområdet. En skole eller et kontor i nærområdet er veldig gunstig med

tanke på at kunder som regel kjøper mat i farten, da er det veldig tilgjengelig å kunne stikke innom butikken etter jobb eller skole, noe som er positivt for butikkens omsetning og profitt. Dette er to faktorer som kan ha en betydning på kundestrømmen, og øke antall solgte enheter som da vil øke profitten. Månedsleien og konkurrerende kjeder i nærområdet har også en betydelig påvirkning på butikkens salg. Vår anbefaling er å åpne en butikk men drive-throug, nær en skole eller kontorer i nærområdet, og helst prøve å unngå å ha mange konkurrerende kjeder i nærområdet.

Referanser:

Trent J. Spaulding, Edgar E. Hassler, Charles H.L. Edwards, Joseph A. Cazier, "Sandwich analytics: A dataset comprising one year's weekly sales data correlated with crime, demographics, and weather", Volume 25, August 2019.