Applying the CRISP-DM Data Science Methodology to Sales Volume Forecasting and Budgeting Problems

Data Understanding

Matthias Hofmaier (11944050)

May 2023

II. Data Understanding Next is the Data Understanding phase. Adding to the foundation of Business Understanding, it drives the focus to identify, collect, and analyze the data sets that can help you accomplish the project goals. This phase also has four tasks:

Collect initial data: Acquire the necessary data and (if necessary) load it into your analysis tool. Describe data: Examine the data and document its surface properties like data format, number of records, or field identities. Explore data: Dig deeper into the data. Query it, visualize it, and identify relationships among the data. Verify data quality: How clean/dirty is the data? Document any quality issues.

Imports

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(modeest)
```

```
## Registered S3 method overwritten by 'rmutil':
##   method         from
##   print.response httr
```

```
library(zoo)
```

```
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(ggplot2)
library(xtable)
```

Constants

```
BASE_PATH <- "../data/processed"
SALES_PATH <- paste(BASE_PATH, "sales.csv", sep = "/")
BALANCE_SHEET_PATH <- paste(BASE_PATH, "balance_sheet.csv", sep = "/")
PROFIT_LOSS_PATH <- paste(BASE_PATH, "profit_loss.csv", sep = "/")
```

Sales

Load data

```
df_sales <- read_csv(SALES_PATH, show_col_types = FALSE)
```

```
## New names:
## * `` -> `...1`
```

```
df_sales <- df_sales[, -1] # remove index column
xtable(head(df_sales), type="html")
```

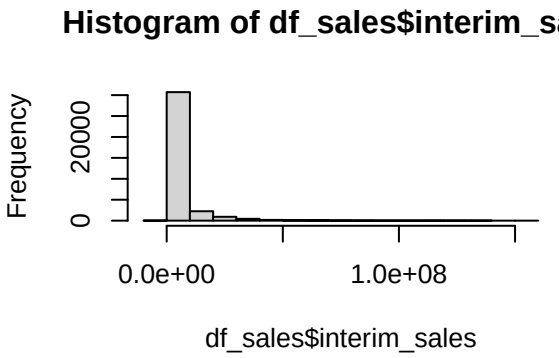
```
## % latex table generated in R 4.2.1 by xtable 1.8-4 package
## % Mon May  1 16:01:26 2023
## \begin{table}[ht]
## \centering
## \begin{tabular}{rlrrr}
##   \hline
##  & company & interim\_sales & year & quarter \\
##   \hline
## 1 & APPLE INC & 1475000.00 & 2003.00 & 1.00 \\
##   2 & APPLE INC & 1909000.00 & 2004.00 & 1.00 \\
##   3 & APPLE INC & 3243000.00 & 2005.00 & 1.00 \\
##   4 & APPLE INC & 4359000.00 & 2006.00 & 1.00 \\
##   5 & APPLE INC & 5264000.00 & 2007.00 & 1.00 \\
##   6 & APPLE INC & 7512000.00 & 2008.00 & 1.00 \\
##    \hline
## \end{tabular}
## \end{table}
```

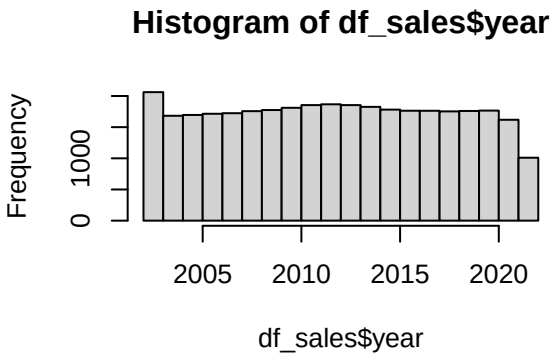
Show data description

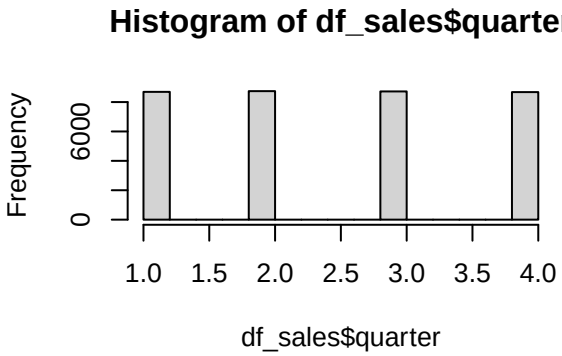
```
##                    type n_distinct        min       max
## company       character        500 3M COMPANY    ZOETIS
## interim_sales   numeric      29190    -393000 152859000
## year            numeric         21       2002      2022
## quarter         numeric          4          1         4
```

Distributions

```
par(mfrow=c(2,2)) # set 2x2 plot grid
hist(df_sales$interim_sales)
hist(df_sales$year)
hist(df_sales$quarter)
```



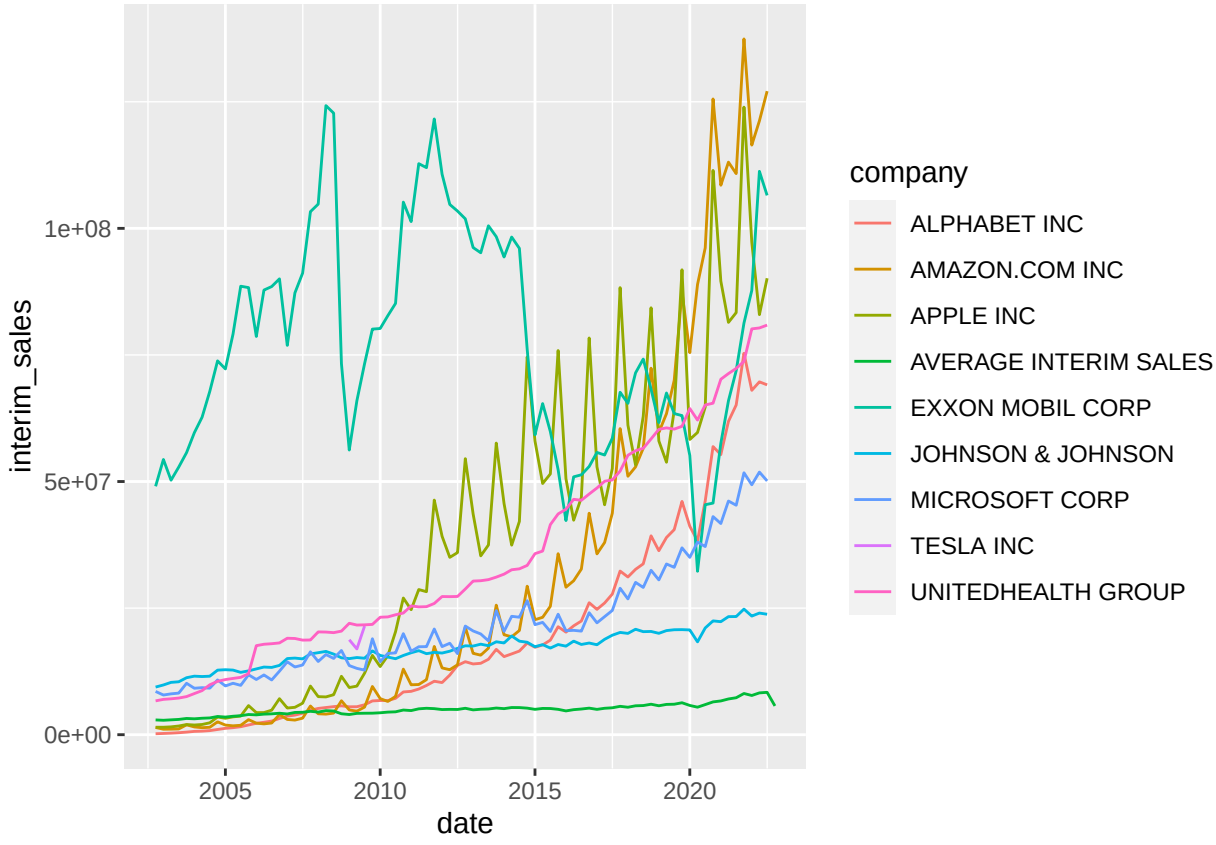




Change over time

```
# create date from year and quarter
df_sales$date <-
  as.Date(as.yearqtr(paste0(df_sales$year, "-", df_sales$quarter), format = "%Y-%q"))

# calculate average over all companies
average_interim_sales <-
  df_sales %>% group_by(date) %>% summarise(interim_sales = mean(interim_sales))
average_interim_sales$company <- "AVERAGE INTERIM SALES"

# define selection of companies as we cannot visualise all companies
selected_companies <-
  c(
    "APPLE INC",
    "MICROSOFT CORP",
    "AMAZON.COM INC",
    "TESLA INC",
    "ALPHABET INC",
    "UNITEDHEALTH GROUP",
    "EXXON MOBIL CORP",
    "JOHNSON & JOHNSON"
  )
df_sales_selected <-
  df_sales[df_sales$company %in% selected_companies,
           c("date", "interim_sales", "company")]

# bind selected companies with average over all companies
```

```r
df_sales_selected <- rbind(df_sales_selected, average_interim_sales)


# show line plot
ggplot(df_sales_selected, aes(x = date, y = interim_sales)) +
  geom_line(aes(color = company))
```



Data quality assesment

```r
data.frame(
  absolute_missing_values = colSums(is.na.data.frame(df_sales)),
  relative_missing_values = colSums(is.na.data.frame(df_sales)) / length(df_sales)
)
```

```
##                absolute_missing_values relative_missing_values
## company                              0                        0
## interim_sales                        0                        0
## year                                 0                        0
## quarter                              0                        0
## date                                 0                        0
```

Balance sheet

Load data

```
df_balance_sheet <- read_csv(BALANCE_SHEET_PATH, show_col_types = FALSE)
```

```
## New names:
## * `` -> `...1`
```

```
df_balance_sheet <- df_balance_sheet[, -1]
head(df_balance_sheet)
```

```
## # A tibble: 6 x 30
##   company  year BORROW~1 EQUIT~2 NET C~3 NET D~4 ORDIN~5 PREFE~6 TOTAL~7 ASSET~8
##   <chr>   <dbl>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 APPLE    2002        0  4.10e6  3.73e6 -4.02e6 1826000       0 2269000  6.23e6
## 2 APPLE    2003   304000  4.22e6  3.53e6 -4.26e6 1926000       0 2297000  6.76e6
## 3 APPLE    2004        0  5.08e6  4.38e6 -5.46e6 2514000       0 2562000  7.96e6
## 4 APPLE    2005        0  7.47e6  6.82e6 -8.26e6 3521000       0 3945000  1.14e7
## 5 APPLE    2006        0  9.98e6  8.04e6 -1.01e7 4355000       0 5629000  1.72e7
## 6 APPLE    2007        0  1.45e7  1.27e7 -1.54e7 5368000       0 9164000  2.53e7
## # ... with 20 more variables: `TOTAL ASSETS EMPLOYED` <dbl>,
## #   `TOTAL CAPITAL EMPLOYED` <dbl>, `TOTAL CASH & EQUIVALENT` <dbl>,
## #   `TOTAL CURRENT ASSETS` <dbl>, `TOTAL CURRENT LIABLITIES` <dbl>,
## #   `TOTAL DEBT` <dbl>, `TOTAL DEBTORS & EQUIVALENT` <dbl>,
## #   `TOTAL DEFERRED & FUTURE TAX` <dbl>, NET <dbl>, `TOTAL INTANGIBLES` <dbl>,
## #   `TOTAL INVESTMNTS (EX.ASSOC)` <dbl>, `TOTAL LOAN CAPITAL` <dbl>,
## #   `TOT. SHARE CAPITAL & RESERVES` <dbl>, `TOTAL STOCK AND W.I.P.` <dbl>, ...
```

Show data description

```
show_data_description(df_balance_sheet)
```

```
##                                    type n_distinct        min        max
## company                       character        503         3M   ZOETIS A
## year                            numeric         21       2002       2022
## BORROWINGS REPAYABLE < 1 YEAR   numeric       6323          0  484315900
## EQUITY CAP. AND RESERVES        numeric       9558  -25560000  506198800
## NET CURRENT ASSETS              numeric       7617 -149782000  290101800
## NET DEBT                        numeric       9485 -173495000  772553000
## ORDINARY SHARE CAPITAL          numeric       4345          0  158142000
## PREFERENCE CAPITAL              numeric        845     -49000   72148000
## TOTAL RESERVES                  numeric       9269  -40796990  506190800
## ASSETS (TOTAL)                  numeric       9685       1893 2119852000
## TOTAL ASSETS EMPLOYED           numeric       9617  -24160000  616640800
## TOTAL CAPITAL EMPLOYED          numeric       9617  -24160000  616640800
## TOTAL CASH & EQUIVALENT         numeric       8663          0  722433800
## TOTAL CURRENT ASSETS            numeric       7777       4693  413188900
## TOTAL CURRENT LIABLITIES        numeric       7714        984  248610000
## TOTAL DEBT                      numeric       8747          0  810758900
## TOTAL DEBTORS & EQUIVALENT      numeric       7369          0  263328000
## TOTAL DEFERRED & FUTURE TAX     numeric       7186  -55032000   89678990
## NET                             numeric       9049          0  259651000
## TOTAL INTANGIBLES               numeric       7963          0  310197000
## TOTAL INVESTMNTS (EX.ASSOC)     numeric       4608  -23979870 1591976000
```

```
## TOTAL LOAN CAPITAL              numeric       8555          0  377137900
## TOT. SHARE CAPITAL & RESERVES   numeric       9550  -25560000  506198800
## TOTAL STOCK AND W.I.P.          numeric       5843          0   81714990
## TRADE CREDITORS                 numeric       6938          0   78664000
## TRADE DEBTORS                   numeric       7808          0  263328000
## TOTAL INSURANCE FUNDS           numeric        497      59565  523148800
## INSURANCE                       numeric       1471          0 1591976000
## CURRENT, DEPOSIT & OTHER A/CS   numeric        567          0 2144257000
## TOTAL ADVANCES                  numeric        631          0 1061328000
```