Big data

Databehandling af data fra et bistade

Armandas Rokas

11/09 2020

Contents

Introduktion

Dette rapport beskriver databehandling af data fra et bistade. Dataen samles via Raspbarry Pi, som er tilknyttet til forskellige sensorer, som måler vægten, fugtighed, lys osv. (hele listen af variabler kan man se i Variabler afsnit), og gemmes både lokalt på SQLite databasen og sendes til HiveTool.net platformen.

Import af data

Der blev valgt at bruge SQLite til at indlæse dataen, da det er lidt svært at få fat i dataen på HiveTool.net, især hvis man skal bruge en lang tidsperiode. Så der blev skrevet en lille funktion, som indlæser dataen fra SQLite, som man kan se nedenfor:

```
import_hive_data <- function(from, to){
  library(DBI)
  con <- dbConnect(RSQLite::SQLite(), "data/stade1.db")
  table_name <- dbListTables(con)
  fields <- dbListFields(con, table_name)
  select_period_with_intervention_query <- paste("SELECT * from", table_name,
                        "WHERE hive_observation_time_local > ",from," AND
                        hive_observation_time_local < ", to, sep=" ")

  sendQuery <- dbSendQuery(con, select_period_with_intervention_query )
  #  hive_data <<- dbFetch(sendQuery)
  hive_data <- dbFetch(sendQuery)
  hive_data$hive_observation_time_local <- strptime(hive_data$hive_observation_time_local,
            format = "%Y-%m-%d %H:%M:%S") #  Convert string to be recognized as date
  return (hive_data)
}
```

Data beskrivelse

Observationer fra bistadet bliver taget hver 5. minut og der er i alt 105064 observationer:

```
nrow(hive_data)
```

```
## [1] 105064
```

Dataen er dog ikke helt konsistent, da der er nogle huller i datasættet, hvor der mangler målinger, og der er nogle målinger, som har forkerte værdier. Dataen går helt tilabge til 2018-03-07, men den bliver mere og mere inkonsistente jo mere i fortiden man går, så derfor blev der valgt at tage udgangspunkt i et års data, dvs. fra 2019-09-01 til 2020-09-01.

Variabler

Der er i alt 43 variabler i datasættet:

```
ncol(hive_data)
```

```
## [1] 43
```

Men ud fra opsummering af alle variabler i datasættet i Bilag A kan man konstatere at der er kun disse variabler er i brug og relevente: `hive_observation_time_local`, `hive_weight_kgs`, `hive_temp_c`, `hive_humidity`, `ambient_temp_c`, `ambient_humidity`, `ambient_luminance`, SPØRGE FREDERIK OMKRING WX data. Nedenfor opsummering af disse variabler:

```
summary(hive_data$hive_observation_time_local)
```

```
##                   Min.                 1st Qu.                  Median
## "2019-09-01 00:00:01" "2019-12-01 06:18:46" "2020-03-01 17:27:31"
##                   Mean                 3rd Qu.                    Max.
## "2020-03-01 17:57:12" "2020-06-01 03:11:16" "2020-08-31 23:55:01"
```

```
summary(hive_data$hive_weight_kgs)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.06   23.65   26.97   28.94   30.36   60.04
```

```
summary(hive_data$hive_temp_c)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.10   20.10   25.00   25.32   34.20   37.10
```

```
summary(hive_data$hive_humidity)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    36.0    57.1    60.3    62.1    67.5    91.0
```

```
summary(hive_data$ambient_temp_c)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -4.300   4.900   8.000   9.496  13.600  31.600    3277
```

```
summary(hive_data$ambient_humidity)
```

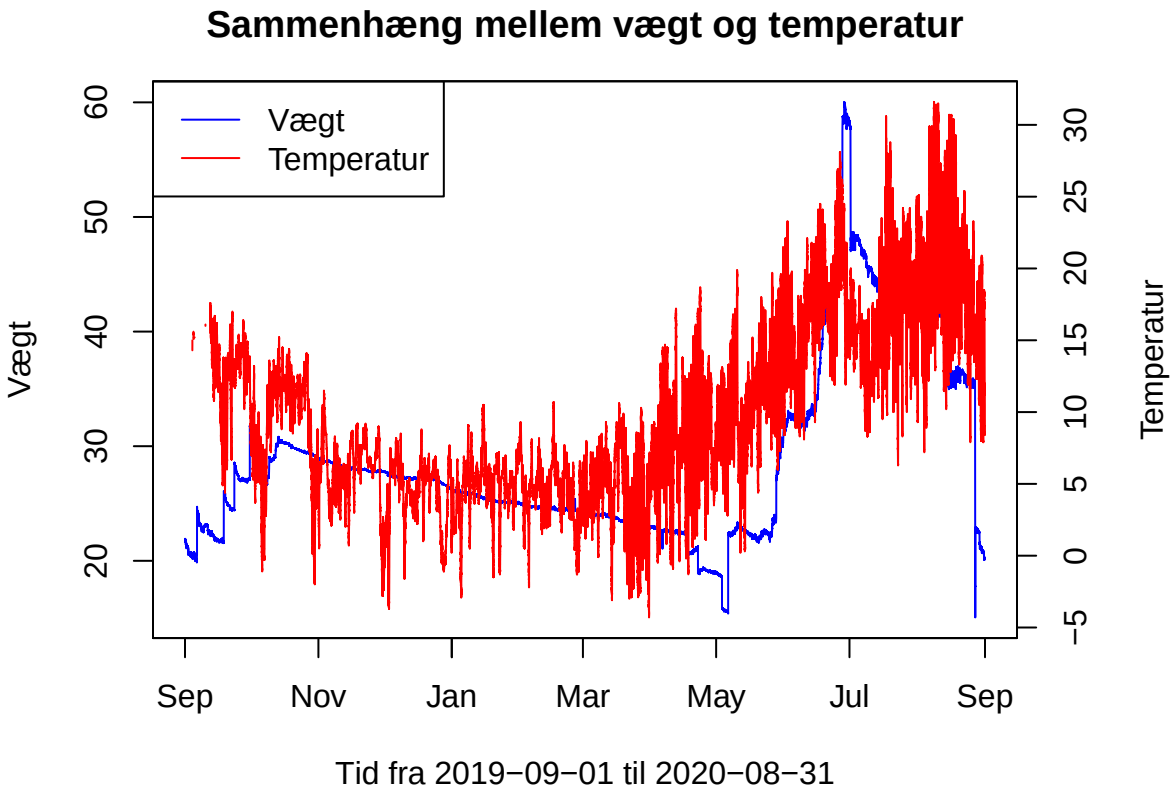
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    55.1    99.9    99.9    99.0    99.9    99.9    3277
```

```
summary(hive_data$ambient_luminance)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   14.65    5.00 2070.00
```

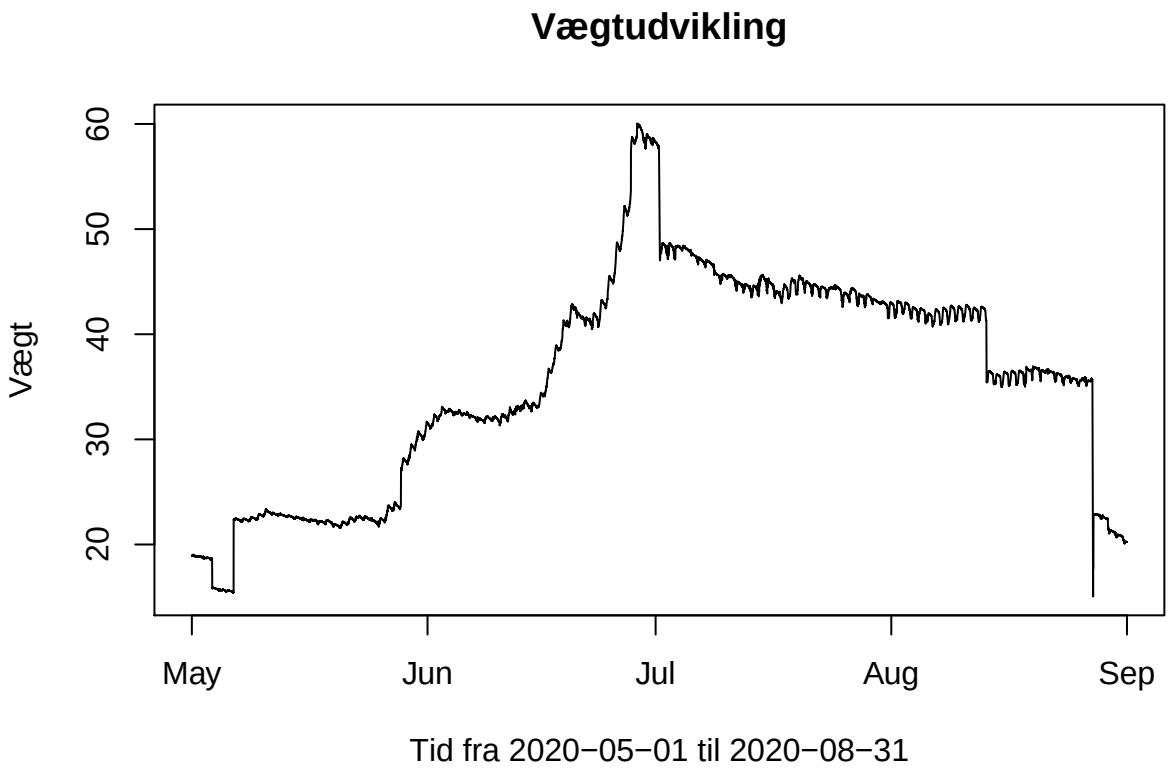
Sammenhæng mellem vægt og temperatur

Ud fra grafen nedenfor kan man fornemme at når tempraturen stiger så vægten gør også, men jeg er ikke sikker om man kan bruge denne sammenhæng mellem vægt og temperatur til noget???? Dvs. giver det overhovet mening til at kigge på correlation i time series??? Linear regression???



Vægten

I dette afsnit fokuseres der kun på vægten, som er den centrale variable i datasættet, i perioden fra 2020-05-01 til 2020-08-31.Dette kan se i grafen nedenfor.



Datarensning Der er ret mange fejlagtig vægtmålinger i datasættet og nedenfor er en liste over mulige grunde, som kan give udsving i vægten, men som ikke forårsagers af bierne:

- Manuelt indgreb
- Nedbør
- ...

Manuelle indgreb Manuelle indgreb på bistadet medfører de største udsving i vægten, så disse skulle fjernes før man kunne påbegynde noget andet. Eksampler på manuelle indgreb kunne være:

- ...

De fleste manual indgreb følge efter afbrydelser i timestamps. Det kan forklares med at en biavler slukker optagelse af målinger, når biavlen laver et manualt indgreb, men når han tænder optagelsen, laver den en stor udsving i vægten i en eller anden retning i for hold hvad han fik lavet.

Så derfor blev der startede med at finde deltaerne mellem vægtene før arbrydelser i timestmaps og efter, dvs. huler i datasættet. En liste over det kan man se nedenfor:

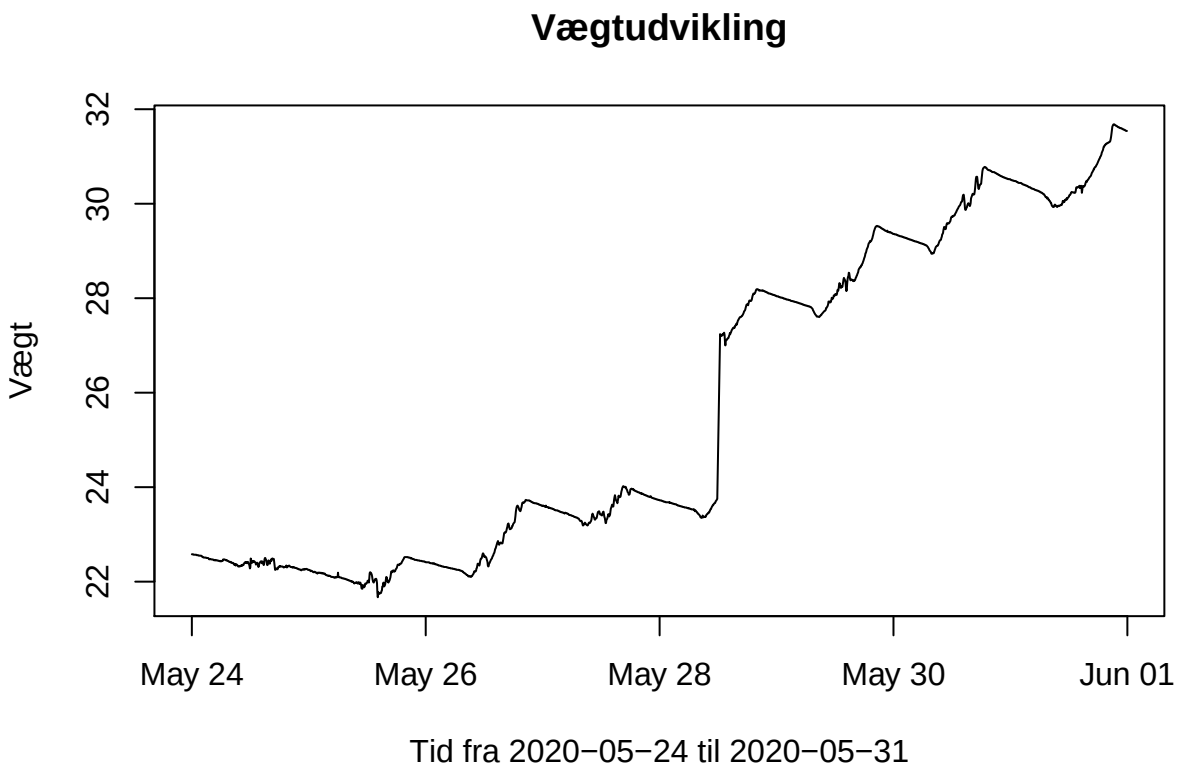
```r
hive_data <- hive_data %>%  mutate(timestamp_delta = hive_observation_time_local -
  dplyr::lag(hive_observation_time_local)) %>%
  mutate(timestamp_delta = ifelse(is.na(timestamp_delta), 0, timestamp_delta))

hive_data <- hive_data %>%  mutate(weight_delta = hive_weight_kgs -
  dplyr::lag(hive_weight_kgs)) %>%
  mutate(weight_delta = ifelse(is.na(weight_delta), 0, weight_delta))

hive_data[which(hive_data[,"timestamp_delta"] > 7),
          c("hive_observation_time_local", "weight_delta")]
```

```
##       hive_observation_time_local weight_delta
## 769           2020-05-03 16:25:01        -2.78
## 1578          2020-05-06 12:00:41         6.90
## 3622          2020-05-13 15:15:01         0.00
## 6810          2020-05-24 17:05:01        -0.23
## 7900          2020-05-28 12:25:01         3.49
## 9025          2020-06-01 11:15:01         0.10
## 11335         2020-06-09 12:55:01        -0.15
## 12795         2020-06-14 15:05:01        -0.34
## 15156         2020-06-22 20:05:01         0.07
## 15668         2020-06-24 15:10:01         0.12
## 15671         2020-06-24 15:45:15        -0.09
## 16563         2020-06-27 18:15:01         3.89
## 17094         2020-06-29 14:50:01        -0.15
## 17341         2020-06-30 11:30:01         0.21
## 17620         2020-07-01 13:30:01       -10.38
## 17851         2020-07-02 08:50:01         0.12
## 19681         2020-07-08 17:45:02         0.21
## 19682         2020-07-08 18:05:35        -0.27
## 21403         2020-07-14 17:45:01         0.04
## 21924         2020-07-16 14:10:01        -0.76
## 22060         2020-07-17 01:35:01        -0.01
## 22267         2020-07-17 19:20:01         0.08
## 22490         2020-07-18 14:00:01         0.03
## 22785         2020-07-19 14:45:01        -0.03
## 22786         2020-07-19 15:00:01        -0.02
## 23083         2020-07-20 15:50:01        -0.02
## 24513         2020-07-25 15:05:02         0.39
## 25132         2020-07-27 18:55:01        -0.12
## 28558         2020-08-08 17:15:55         0.21
## 29937         2020-08-13 12:50:01        -5.90
## 30821         2020-08-16 15:00:01         0.16
## 32396         2020-08-22 02:20:01         0.00
## 32633         2020-08-22 22:10:01         0.00
## 32705         2020-08-23 04:15:01         0.01
## 33937         2020-08-27 13:05:02       -20.61
## 33938         2020-08-27 13:15:56         2.74
## 33948         2020-08-27 14:15:01         5.08
## 34493         2020-08-29 12:25:01        -0.96
```

- Der blev valgt at undersøge vægten videre omkring `2020-05-28 12:25:01`, hvor vægten har øget med `3.49`. Nedenfor er grafen over vægten omkring dette tidspunkt.



Bilag

Bilag A: Opsummering af alle variabler i datasættet

```
summary(hive_data)
```

```
##      rowid           hive_id   timestamp         hive_observation_time_local
##  Min.   : 32730   Min.   :1   Length:105064      Min.   :2019-09-01 00:00:01
##  1st Qu.: 58996   1st Qu.:1   Class :character   1st Qu.:2019-12-01 06:18:46
##  Median : 85262   Median :1   Mode  :character   Median :2020-03-01 17:27:31
##  Mean   : 85262   Mean   :1                      Mean   :2020-03-01 17:57:12
##  3rd Qu.:111527   3rd Qu.:1                      3rd Qu.:2020-06-01 03:11:16
##  Max.   :137793   Max.   :1                      Max.   :2020-08-31 23:55:01
##
##  hive_observation_time_utc hive_weight_lbs  hive_weight_lbs_delta
##  Min.   : NA               Min.   : 33.20   Min.   : NA
##  1st Qu.: NA               1st Qu.: 52.14   1st Qu.: NA
##  Median : NA               Median : 59.46   Median : NA
##  Mean   :NaN               Mean   : 63.81   Mean   :NaN
##  3rd Qu.: NA               3rd Qu.: 66.93   3rd Qu.: NA
##  Max.   : NA               Max.   :132.37   Max.   : NA
##  NA's   :105064                             NA's   :105064
##  hive_weight_lbs_delta_daily hive_weight_lbs_filtered
```

```
##  Min.   : NA                 Min.   : NA
##  1st Qu.: NA                 1st Qu.: NA
##  Median : NA                 Median : NA
##  Mean   :NaN                 Mean   :NaN
##  3rd Qu.: NA                 3rd Qu.: NA
##  Max.   : NA                 Max.   : NA
##  NA's   :105064              NA's   :105064
##  hive_manipulation_change_lbs hive_weight_kgs hive_weight_kgs_delta
##  Min.   : NA                  Min.   :15.06   Min.   : NA
##  1st Qu.: NA                  1st Qu.:23.65   1st Qu.: NA
##  Median : NA                  Median :26.97   Median : NA
##  Mean   :NaN                  Mean   :28.94   Mean   :NaN
##  3rd Qu.: NA                  3rd Qu.:30.36   3rd Qu.: NA
##  Max.   : NA                  Max.   :60.04   Max.   : NA
##  NA's   :105064                               NA's   :105064
##  hive_weight_kgs_delta_daily hive_weight_kgs_filtered
##  Min.   : NA                 Min.   : NA
##  1st Qu.: NA                 1st Qu.: NA
##  Median : NA                 Median : NA
##  Mean   :NaN                 Mean   :NaN
##  3rd Qu.: NA                 3rd Qu.: NA
##  Max.   : NA                 Max.   : NA
##  NA's   :105064              NA's   :105064
##  hive_manipulation_change_kgs  hive_temp_f     hive_temp_c    hive_humidity
##  Min.   : NA                   Min.   :35.78   Min.   : 2.10   Min.   :36.0
##  1st Qu.: NA                   1st Qu.:68.18   1st Qu.:20.10   1st Qu.:57.1
##  Median : NA                   Median :77.00   Median :25.00   Median :60.3
##  Mean   :NaN                   Mean   :77.57   Mean   :25.32   Mean   :62.1
##  3rd Qu.: NA                   3rd Qu.:93.56   3rd Qu.:34.20   3rd Qu.:67.5
##  Max.   : NA                   Max.   :98.78   Max.   :37.10   Max.   :91.0
##  NA's   :105064
##  hive_battery_voltage ambient_temp_f  ambient_temp_c   ambient_humidity
##  Min.   :2.13         Min.   :24.26   Min.   :-4.300   Min.   :55.1
##  1st Qu.:2.13         1st Qu.:40.82   1st Qu.: 4.900   1st Qu.:99.9
##  Median :2.13         Median :46.40   Median : 8.000   Median :99.9
##  Mean   :2.13         Mean   :49.09   Mean   : 9.496   Mean   :99.0
##  3rd Qu.:2.13         3rd Qu.:56.48   3rd Qu.:13.600   3rd Qu.:99.9
##  Max.   :2.13         Max.   :88.88   Max.   :31.600   Max.   :99.9
##                       NA's   :3277    NA's   :3277     NA's   :3277
##  ambient_luminance ambient_precip_in wx_station_id
##  Min.   :   0.00   Min.   : NA       Length:105064
##  1st Qu.:   0.00   1st Qu.: NA       Class :character
##  Median :   0.00   Median : NA       Mode  :character
##  Mean   :  14.65   Mean   :NaN
##  3rd Qu.:   5.00   3rd Qu.: NA
##  Max.   :2070.00   Max.   : NA
##                    NA's   :105064
##  wx_observation_time_rfc822   wx_temp_f        wx_temp_c
##  Min.   : NA                Min.   : 21.60   Min.   :  0.00
##  1st Qu.: NA                1st Qu.: 41.00   1st Qu.:  5.00
##  Median : NA                Median : 45.30   Median :  7.40
##  Mean   :NaN                Mean   : 48.52   Mean   :  9.41
##  3rd Qu.: NA                3rd Qu.: 53.60   3rd Qu.: 12.00
##  Max.   : NA                Max.   :999.00   Max.   :572.80
```

```
##  NA's   :105064             NA's   :69201    NA's   :69201
##  wx_relative_humidity wx_wind_dir        wx_wind_degrees   wx_wind_mph
##  Min.   : 46.00       Length:105064      Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 88.00       Class :character   1st Qu.:  0.00   1st Qu.:  0.00
##  Median : 94.00       Mode  :character   Median :  0.00   Median :  0.20
##  Mean   : 93.14                          Mean   : 38.59   Mean   :  2.84
##  3rd Qu.: 98.00                          3rd Qu.: 37.00   3rd Qu.:  1.30
##  Max.   :999.00                          Max.   :999.00   Max.   :999.00
##  NA's   :69201                                            NA's   :69201
##  wx_wind_gust_mph wx_pressure_mb   wx_pressure_in   wx_dewpoint_f
##  Min.   :  0.00   Min.   : 973.5   Min.   : NA      Min.   : 20.80
##  1st Qu.:  0.00   1st Qu.: 995.8   1st Qu.: NA      1st Qu.: 38.80
##  Median :  1.10   Median :1003.3   Median : NA      Median : 43.30
##  Mean   :  3.36   Mean   :1002.4   Mean   :NaN      Mean   : 45.96
##  3rd Qu.:  2.20   3rd Qu.:1009.7   3rd Qu.: NA      3rd Qu.: 50.70
##  Max.   :999.00   Max.   :1035.8   Max.   : NA      Max.   :999.00
##  NA's   :69201    NA's   :69201    NA's   :105064   NA's   :69201
##  wx_dewpoint_c    wx_solar_radiation wx_precip_1hr_in wx_precip_1hr_metric
##  Min.   :  0.00   Min.   : NA        Min.   :  0.00   Min.   : NA
##  1st Qu.:  3.90   1st Qu.: NA        1st Qu.:  0.00   1st Qu.: NA
##  Median :  6.30   Median : NA        Median :  0.00   Median : NA
##  Mean   :  8.08   Mean   :NaN        Mean   :  1.95   Mean   :NaN
##  3rd Qu.: 10.40   3rd Qu.: NA        3rd Qu.:  0.00   3rd Qu.: NA
##  Max.   :572.80   Max.   : NA        Max.   :999.00   Max.   : NA
##  NA's   :69201    NA's   :105064     NA's   :69201    NA's   :105064
##  wx_precip_today_in wx_precip_today_metric    quality
##  Min.   :  0.00     Min.   : NA            Min.   :5
##  1st Qu.:  0.00     1st Qu.: NA            1st Qu.:5
##  Median :  0.00     Median : NA            Median :5
##  Mean   :  2.00     Mean   :NaN            Mean   :5
##  3rd Qu.:  0.03     3rd Qu.: NA            3rd Qu.:5
##  Max.   :999.00     Max.   : NA            Max.   :5
##  NA's   :69201      NA's   :105064
```