# DATASET_REPORT

*Ingenieros del Futuro*

*20 de octubre de 2019*

## Loading libraries

```
library(tidyverse)
```

```
## -- Attaching packages -------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.0     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

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

```
library(readxl)
library(lubridate)
```

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

## The following object is masked from 'package:base':
##
##     date
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.6.1

## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(stats)
```

## Dataset location

```
dataset_raw = read.csv("ldex_20161118/data_calibrated/mass/DATASET.tab",
                       sep = " ")
```

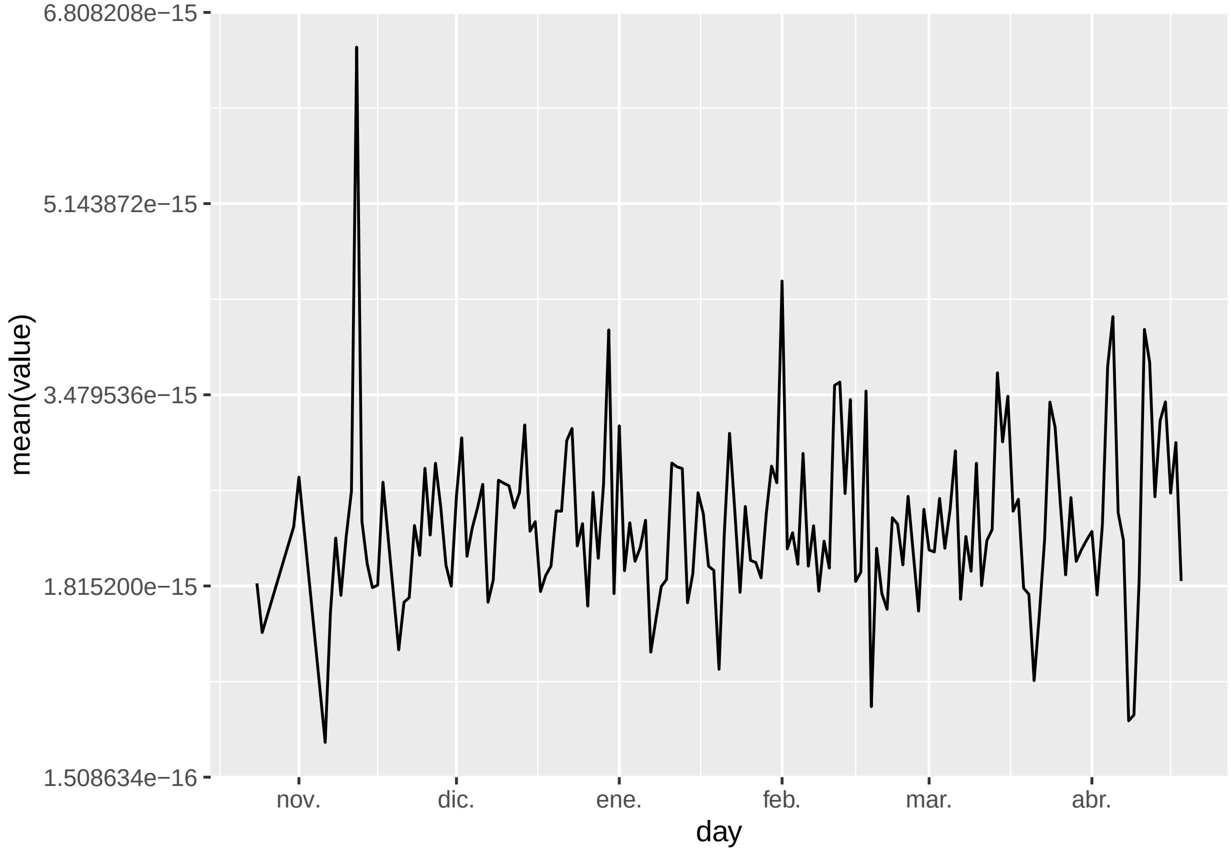
## Tidy data

The next step was to tidy the dataset. For the extension of code, this would not be included but is located in the RScript_2.R

## Data Plot

For data analysis we would make a plot. In this plot we observe a cyclical behavior of the data

```
sum_dat %>%
  ggplot(aes(day, `mean(value)`)) +
  geom_line()
```



## Fit a model prediction

Next step was to fit a prediction model for the dataset in order to be able to predict future or not given data points

```
p <- 0.8
set.seed(1)

test_index = sample.int(n = nrow(sum_dat),
                        size = floor(p*nrow(sum_dat)),
                        replace = FALSE)
train = sum_dat[test_index,]
test = sum_dat[-test_index,]

model = lm(`mean(value)` ~ day, data = train)
valuepred = predict(model, test)

summary(model)
```

```
##
## Call:
## lm(formula = `mean(value)` ~ day, data = train)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.772e-15 -3.957e-16 -6.032e-17  3.565e-16  1.751e-15
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.418e-14  1.744e-14  -1.386    0.168
## day          1.645e-18  1.084e-18   1.518    0.132
##
## Residual standard error: 6.148e-16 on 130 degrees of freedom
## Multiple R-squared:  0.01741,    Adjusted R-squared:  0.009851
## F-statistic: 2.303 on 1 and 130 DF,  p-value: 0.1315
```

```
actual_preds <- data.frame(cbind(actuals = test$`mean(value)`,
                                 predicted = valuepred))
actual_preds['error'] = actual_preds$actuals - actual_preds$predicted
actual_preds %>%
  summarize(RMSE = (sum(error)^2)/n())
```

```
##           RMSE
## 1 3.179408e-30
```

We conclude that datapoints are insufficient to fit a prediction model to the dataset. Nevertheless due to the cyclical behaviour we suggest a Exponential Smoothing algorithm with 3 year data collection