194.044 Data Stewardship

Assignment 1

Aleksandar Pavlovic

16.04.2019

Question

This document deals with answering the pseudo question, if the global temperature trend influences the number of children born alive in Vienna.

Datasets

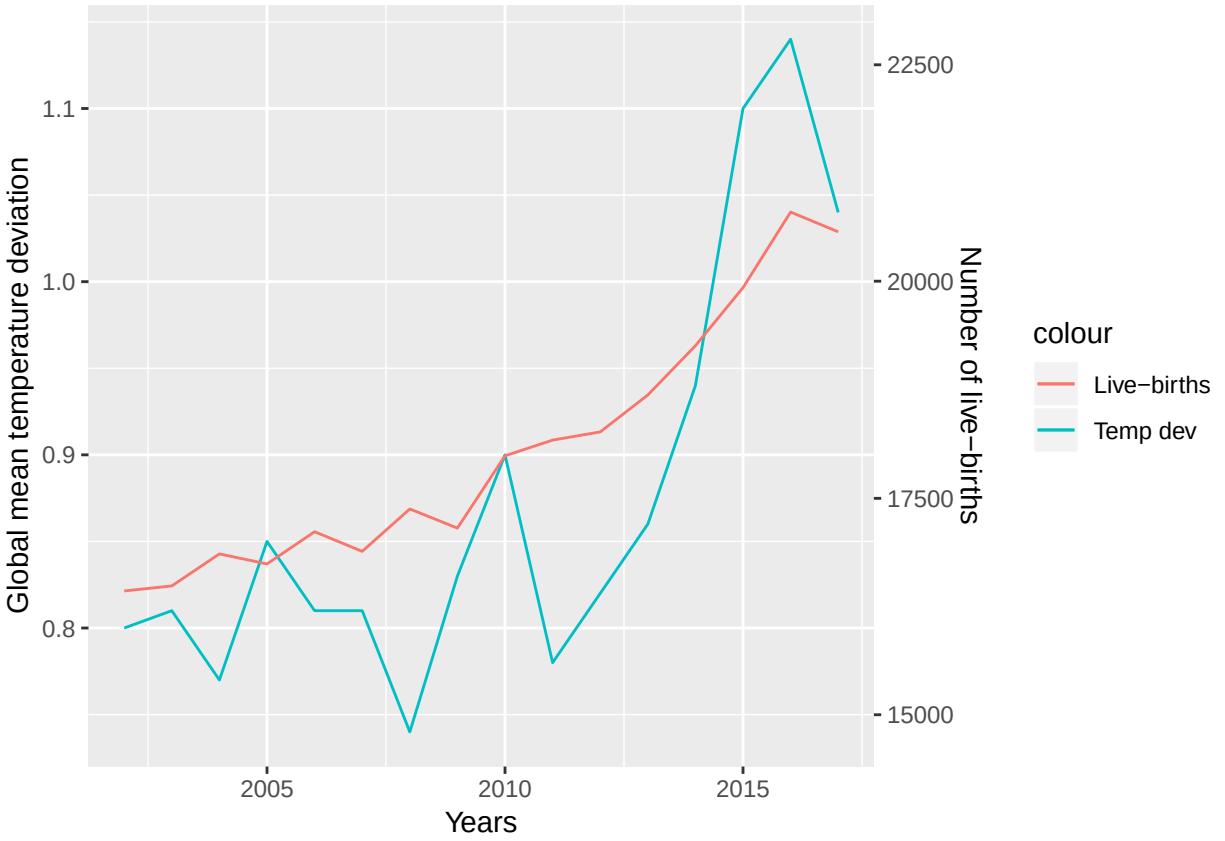
To answer the proposed question above, the following two datasets are used:

- Live births in Vienna - identifier: f54e6828-3d75-4a82-89cb-23c58057bad4, url: [birth data repo](https://www.data.gv.at/katalog/dataset/f54e6828-3d75-4a82-89cb-23c58057bad4)
- Mean near surface temperature deviation (the first zip file was downloaded) - identifier: *cli_iad_td*, url: [temperature repo](https://data.europa.eu/euodp/en/data/dataset/zQAEvhkR7H0NQYU1HP5fA)

```
# Load libraries
library("ggplot2")
```

```
# Load preprocessed and merged data
data <- read.csv2("./data/processed/mergedData.csv")
```

```
# First visualisation:
ggplot(data=data, aes(x=year, group=1)) +
  labs(y = "Global mean temperature deviation",
                x = "Years") +
  geom_line(aes(y = GlobalTempDev, colour = "Temp dev")) +
  geom_line(aes(y = LIVEBIRTH/20000, colour = "Live-births")) +
  scale_y_continuous(sec.axis = sec_axis(~.*20000, name = "Number of live-births"))
```

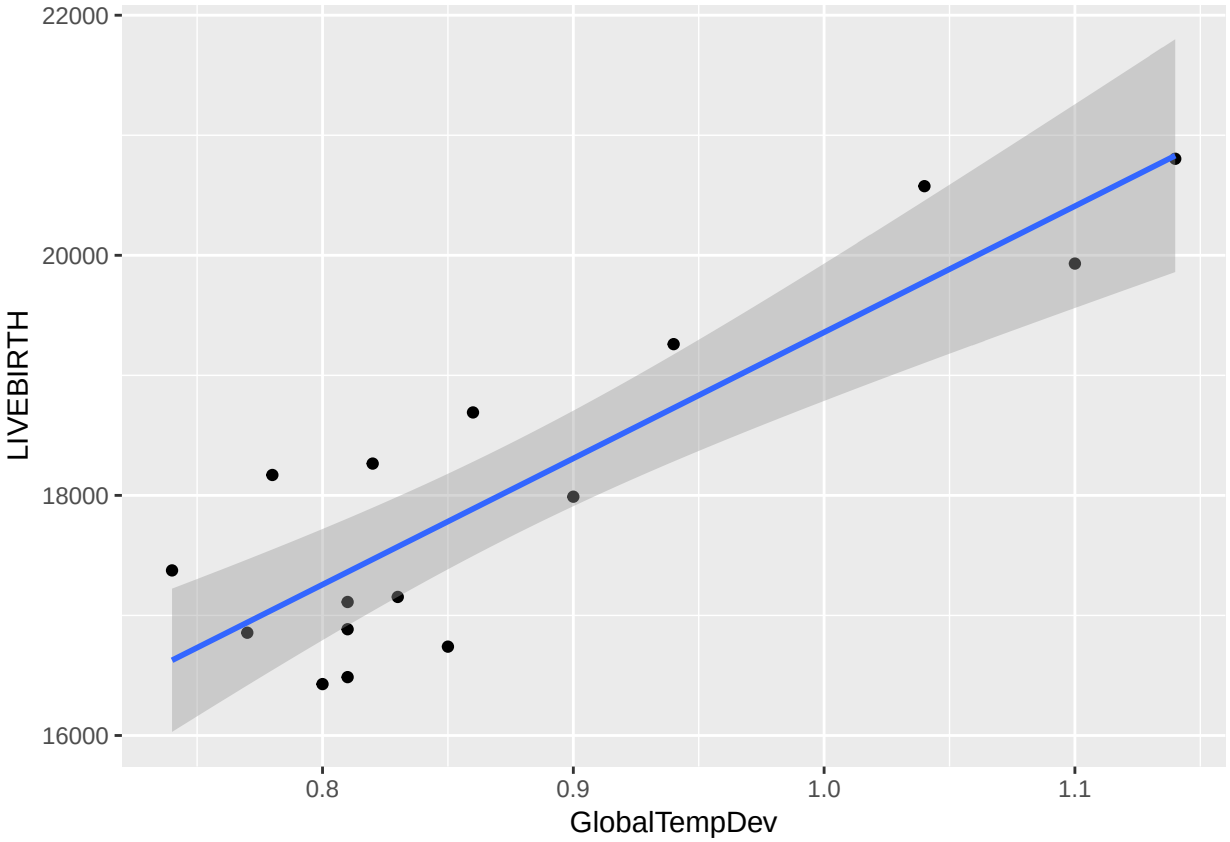



```
ggsave("./plots/timeSeriesLinePlot.jpg")
```

```
## Saving 6.5 x 4.5 in image
```

One can see in the graph above, that the general trend of both variables is the same (thus the two variables seem to be highly correlated).

```
# Second visualisation:
ggplot(data=data, aes(x=GlobalTempDev, group=1)) +
  geom_point(aes(y = LIVEBIRTH)) +
  geom_smooth(mapping = aes(x = GlobalTempDev, y = LIVEBIRTH), method=lm)
```




```
ggsave("./plots/scatterPlotWithLinearModel.jpg")
```

```
## Saving 6.5 x 4.5 in image
```

Further investigation leads to the above scatterplot enhanced with the prediction of a linear model (blue line) and the 95% confidence interval (grey area). One can see, that all black points (data points) are very close to that prediction line, thus again revealing, that the two variables are highly correlated.

```
cor(data$GlobalTempDev, data$LIVEBIRTH, method="pearson")
```

```
## [1] 0.8738113
```

Looking at the Pearson correlation coefficient reveals, that the data is in fact really highly correlated, thus when one ignores that there is a difference between causality and correlation, one can infer that the rising temperature leads to an increase of live-births in Vienna.