07-normalisation-regression.Rmd

Normalising data and visualising trends

Dmytro Fishman

12 April 2016

Authors: **Dmytro Fishman**

Learning Objectives

- Understand the need and idea of data normalisation
- Fit the regression line to the data from global climate trends
- Draw the figure

Load required packages

```
# plotting package
library(ggplot2)
# piping / chaining
library(magrittr)
# modern dataframe manipulations
library(dplyr)
```

```
#>
#> Attaching package: 'dplyr'
#>
#> The following objects are masked from 'package:stats':
#>
#>     filter, lag
#>
#> The following objects are masked from 'package:base':
#>
#>     intersect, setdiff, setequal, union
```

```
temperature_raw <- read.csv('data/temperature.csv')
```

Preprocess data

```
temperature_complete <- temperature_raw %>%
                    filter(!is.na(City)) %>%
                    filter(!is.na(Country)) %>%
                    filter(!is.na(AverageTemperatureFahr)) %>%
                    filter(!is.na(AverageTemperatureUncertaintyFahr))
temperature_complete <- select(temperature_complete, -(day))

temperature_complete <- temperature_complete %>%
  mutate(AverageTemperatureCelsius = (AverageTemperatureFahr-32)*(5/9)) %>%
  mutate(AverageTemperatureUncertaintyCelsius = (AverageTemperatureUncertaintyFahr-32)*(5/9))
```

```
temperature_complete$AverageTemperatureFahr <- NULL
temperature_complete$AverageTemperatureUncertaintyFahr <- NULL
head(temperature_complete)
```

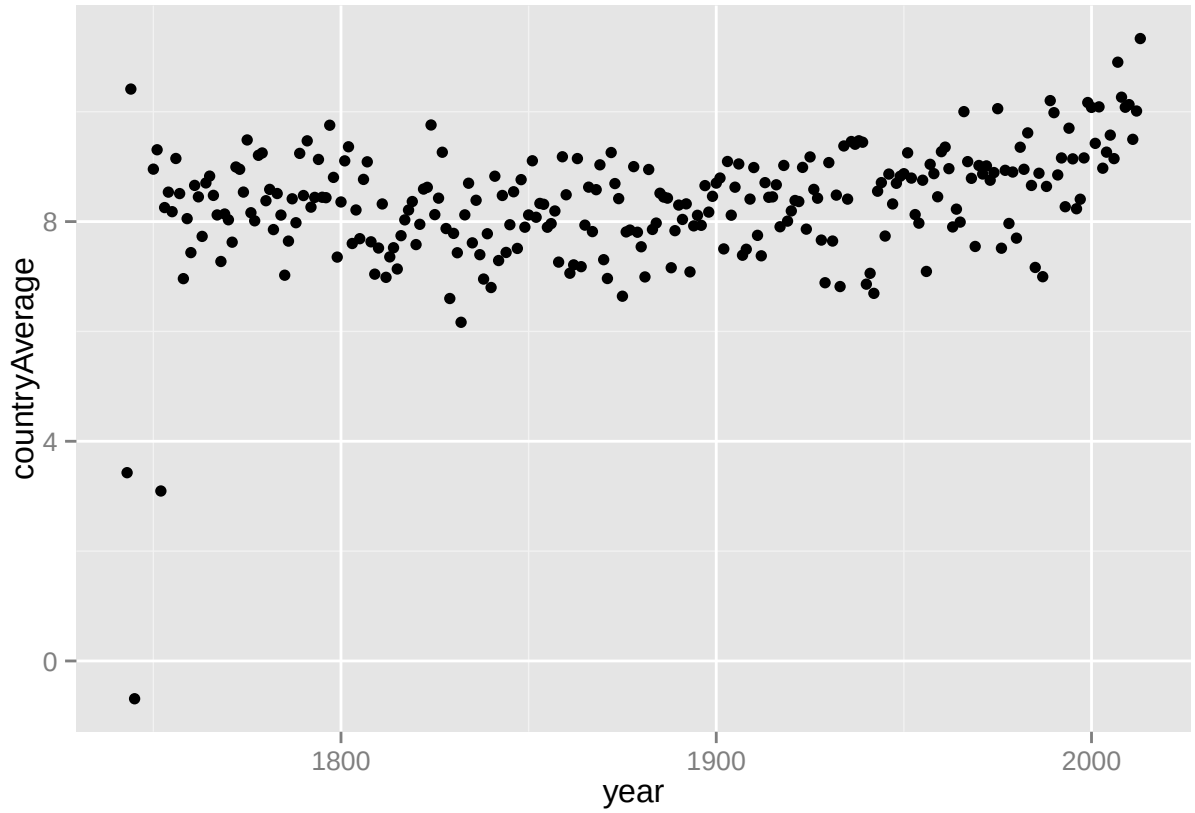
Let's us focus on Ukrainian temperature

```
temperature_ukraine  <- temperature_complete %>%
                        filter(Country == 'Ukraine')
```

Let's visualise the temperature progress per year for this we need to ...

```
yearly_ukraine_temp <- temperature_ukraine %>%
                        group_by(year) %>%
                        summarise(countryAverage = mean(AverageTemperatureCelsius))
```

```
ggplot(data = yearly_ukraine_temp, aes(x = year, y = countryAverage)) +
      geom_point()
```



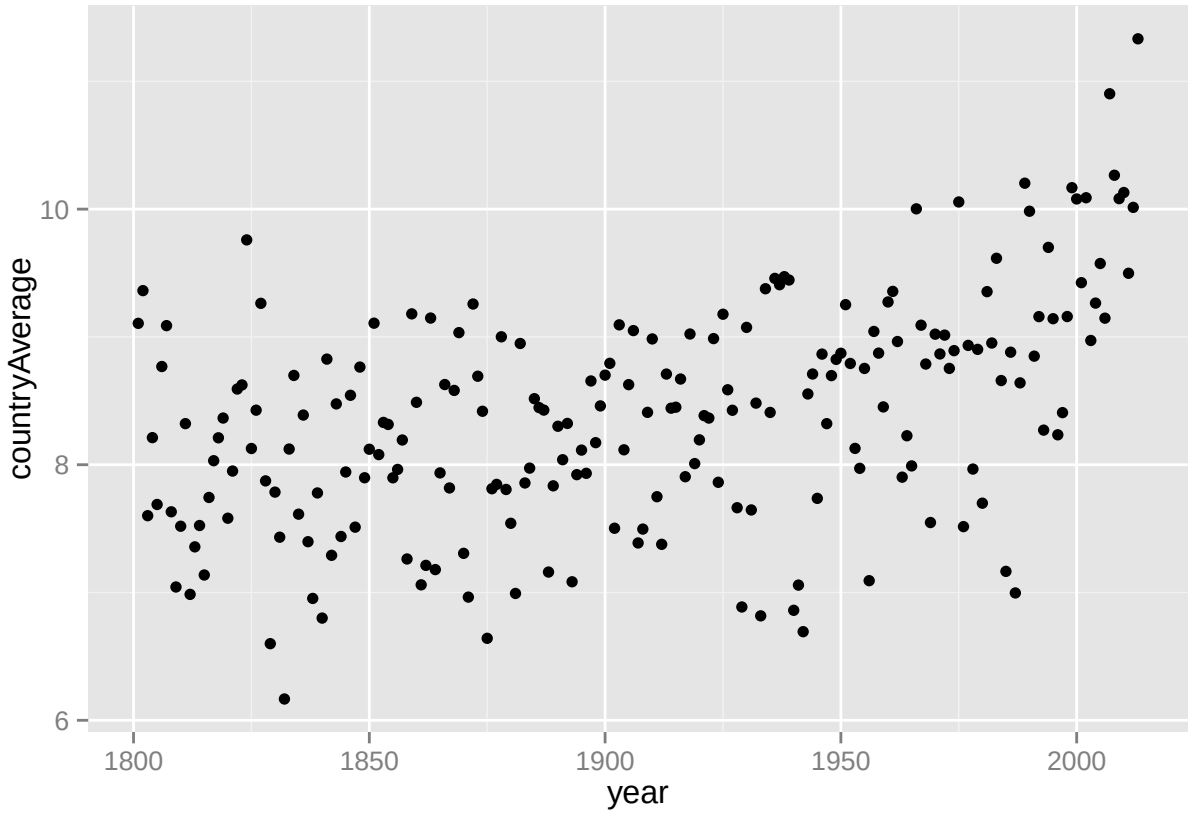
It seems that we don't have enough data until 1800, let's filter it

```
temperature_ukraine <- temperature_ukraine %>%
                        filter(year > 1800)
```

... and try visualising the same information again

```
yearly_ukraine_temp <- temperature_ukraine %>%
                       group_by(year) %>%
                       summarise(countryAverage = mean(AverageTemperatureCelsius))
```

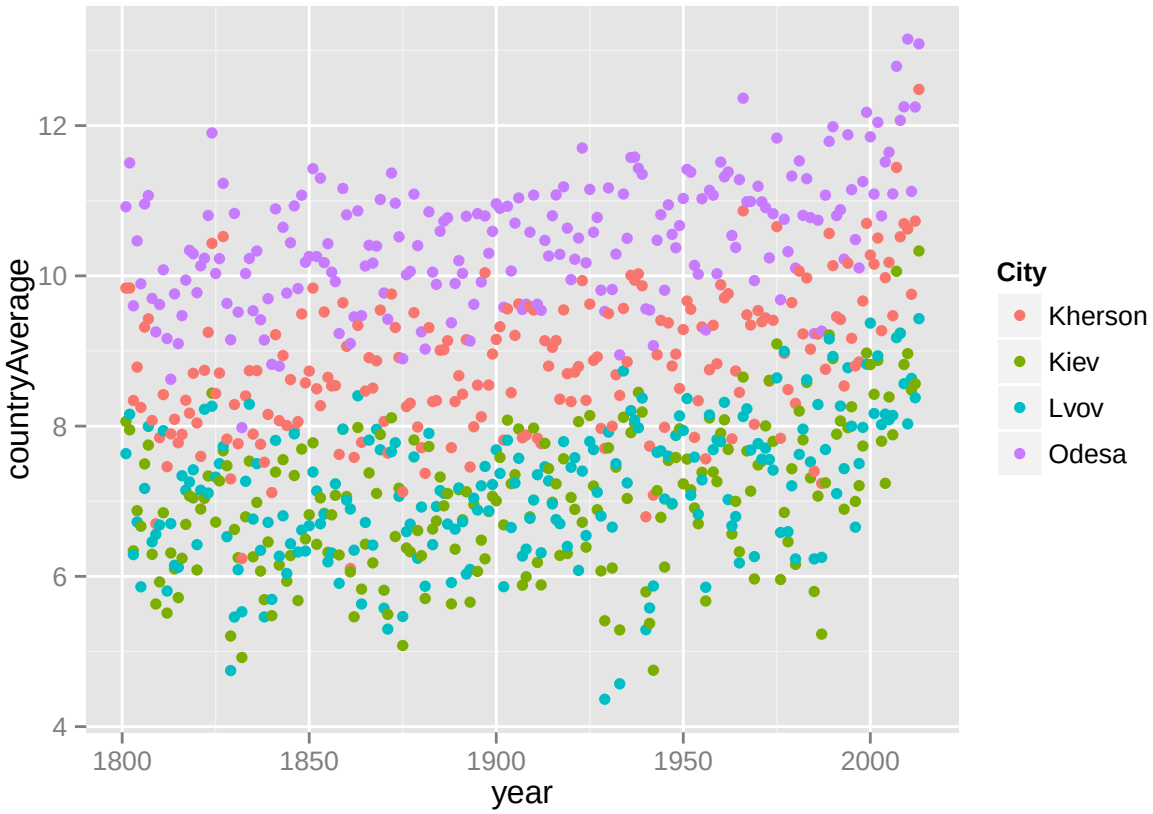
```
ggplot(data = yearly_ukraine_temp, aes(x = year, y = countryAverage)) +
      geom_point()
```



The interesting question here is how much each city influences this trend?

```
yearly_ukraine_temp <- temperature_ukraine %>%
                       group_by(year, City) %>%
                       summarise(countryAverage = mean(AverageTemperatureCelsius))
```

```
ggplot(data = yearly_ukraine_temp, aes(x = year, y = countryAverage, group = City, colour = City)) +
      geom_point()
```



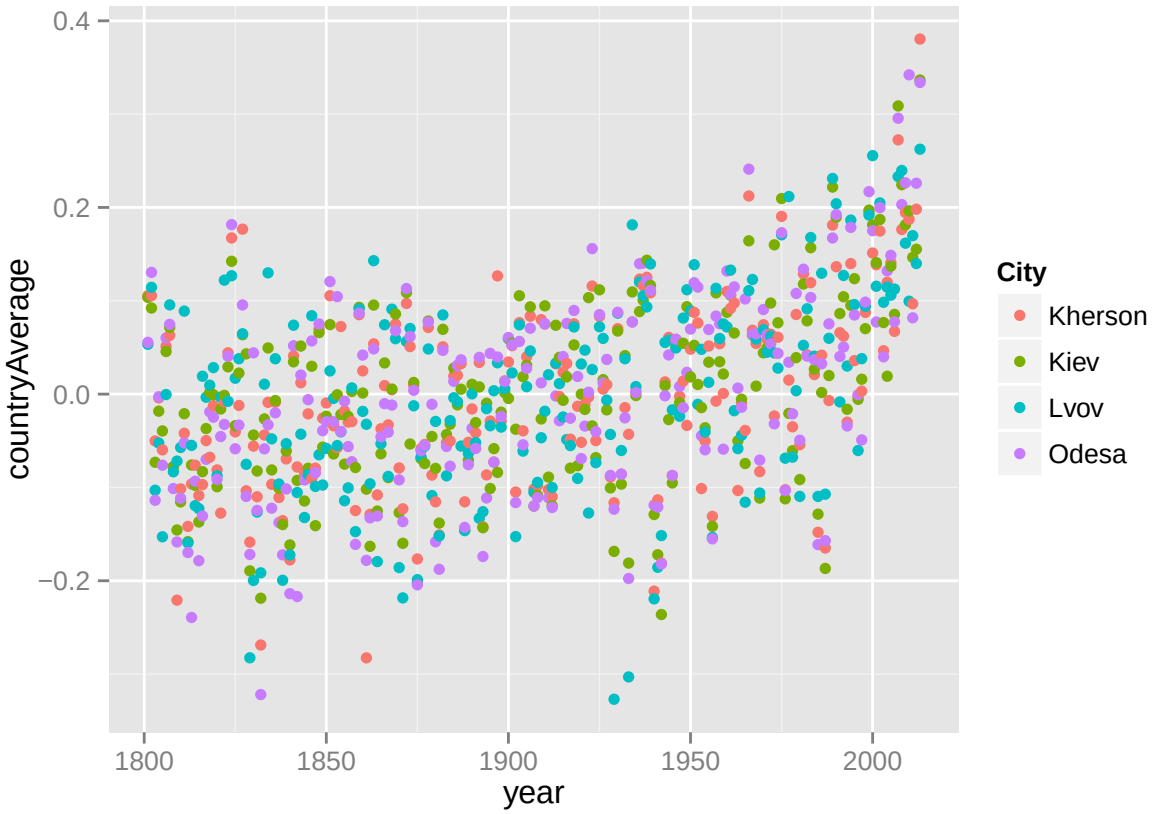
Let's scale all the distributions by average of each city and divide by the standard deviation of each city For that we need to ...

```
scaled_temperature_ukraine <- temperature_ukraine %>%
                              group_by(City) %>%
                              mutate(scaledTemperature = scale(AverageTemperatureCelsius)[,1])
```

Now let's plot the same

```
yearly_ukraine_temp <- scaled_temperature_ukraine %>%
                       group_by(year, City) %>%
                       summarise(countryAverage = mean(scaledTemperature))
```

```
ggplot(data = yearly_ukraine_temp, aes(x = year, y = countryAverage, group = City, colour = City)) +
      geom_point()
```

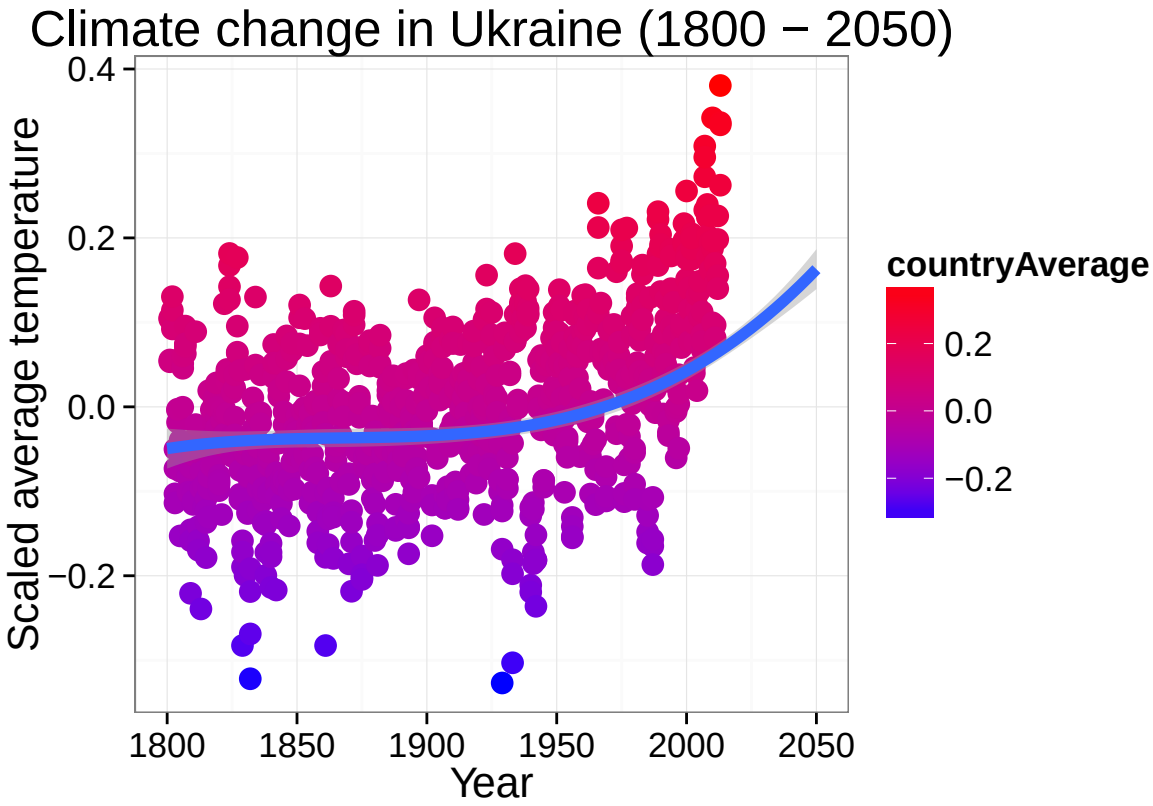


This looks much better, I mean much worse in terms of global warming of course, as it seems that the trend is indeed global.

Regression

Let's add a regression line to the plot above, so that we can see where is the line that goes ....

```
ggplot(data = yearly_ukraine_temp, aes(x = year,  y = countryAverage)) +
      geom_point(aes(colour = countryAverage), size = 4) +
      xlim(1800, 2050) +
      geom_smooth(size = 2, method = lm, formula = y ~ splines::bs(x, 3), fullrange = TRUE) +
      labs(title = 'Climate change in Ukraine (1800 - 2050)',
                        x = 'Year',
                        y = 'Scaled average temperature') +
      theme_bw(base_size = 16) +
      scale_colour_gradient(low = "blue", high = "red")
```



At this level, expected within 40 years, the hot European summer of 2003 will be the annual norm. Anything that could be called a heatwave thereafter will be of Saharan intensity. Even in average years, people will die of heat stress. Global warming, our future