Homework 3

Tim van Doorne

2022-10-20

```
knitr::opts_chunk$set(warning = TRUE, message = TRUE)
library(gapminder)
library(ggplot2)
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
```

```
library(readr)
library(RColorBrewer)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --

## v tibble  3.1.8     v stringr 1.4.1
## v tidyr   1.2.1     v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
io_income_rs <- read_csv("io_income_rs.csv")
```

```
## Rows: 4500 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): donor, type_donor, issue_area
## dbl (2): year, amount_nominal
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Question 1

Create a bar plot displaying the top 10 overall donors and their total donations to all international organizations.Please color the donors by their type (i.e. public or private).

Answer

- Explanation of the steps to wrangle the data:

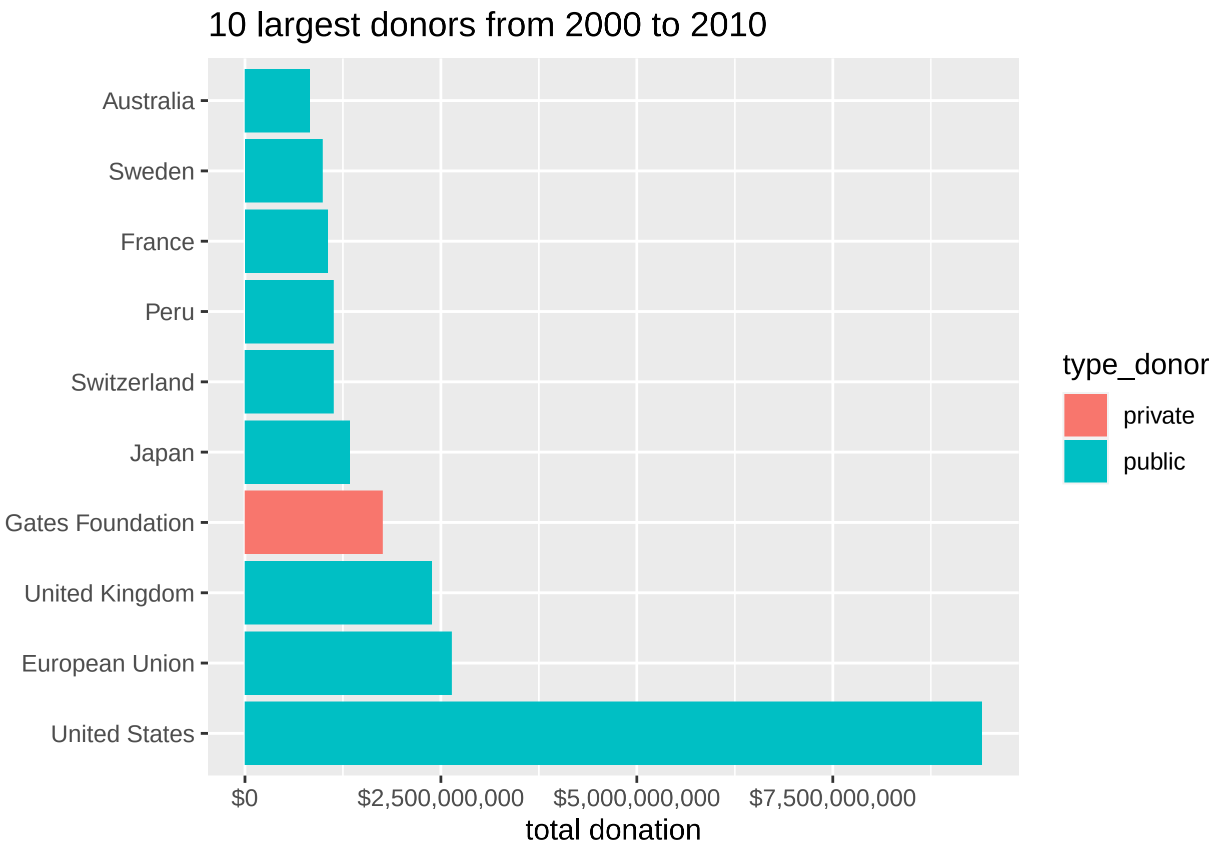
1) drop unknown donors
2) group by donor and donor type
3) for each donor, sum all donor amounts into a total donation, while removing donations with NAs
4) arrange the list of donors based on highest to lowest total donation
5) ungroup by donor and donor type, or else the slice function won't work
6) slice for the top 10 donors with the highest total donation

- Creating the bar plot

1) create a mapping with x-as as 'donor' (reorder donors from lowest to highest), y-as as 'total donation'
2) mapping variables to aesthetics: scale 'total donation' as continuous variable including $ as scale, theme to grey, add labels, flip the bar plot

```
io_income_rs %>%
  drop_na(donor) %>%
  group_by(donor, type_donor) %>%
  summarise(total_donation = sum(amount_nominal, na.rm=TRUE)) %>%
  arrange(-total_donation) %>%
    ungroup() %>% #we need to ungroup, since slice won't work well with grouping
    slice(1:10) %>%
  ggplot(mapping = aes(x=reorder(donor, -total_donation), y=total_donation))+geom_col(mapping=aes(fill=
```

```
## `summarise()` has grouped output by 'donor'. You can override using the
## `.groups` argument.
```



Question 2

Has the amount the US donates to all issue areas increased over time? Please illustrate this relationship in a line plot, colored by the respective issue areas.

Answer

- Wrangling the data:

1) Drop NA donors
2) filter donor for the United States
3) group by year as well as the area of the donation issue
4) for each donation issue, sum all donor amounts into a total donation per year, while removing donations with NAs

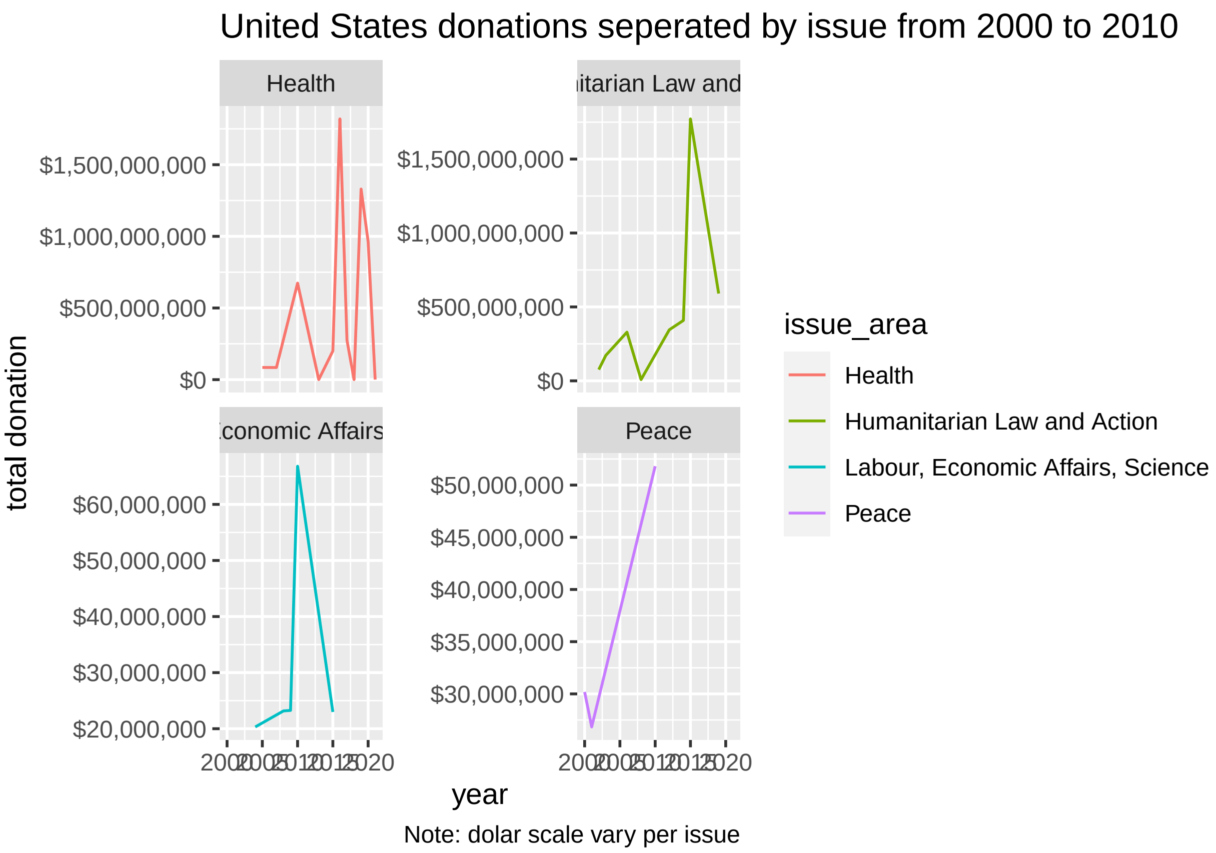
- Creating the line plot

1) Many similar mapping and aesthetics steps from question 1
2) Create different colors for each issue area
3) Facet wrap creates separate line plots for donation issue & gives the plots the freedom to use different $ dollars

- Interpretation: No, the amount of US donates has not increased for each issue area. In fact, donations in Economic Affairs and Peace have even been completely stalled.

```
io_income_rs %>%
  drop_na(donor) %>%
  filter(donor == "United States") %>%
  group_by (year, issue_area) %>%
  summarise (total_donation = sum(amount_nominal, na.rm=TRUE)) %>%
  ggplot(mapping = aes(x=year, y=total_donation, color=issue_area)) + geom_line() +
  facet_wrap("issue_area") + theme_grey() + scale_y_continuous(labels = scales::dollar) + labs(y="total
```

```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```


United States donations seperated by issue from 2000 to 2010

Note: dolar scale vary per issue

Question 3

Plot the distribution of all donations in the year 2000 and in the year 2020 comparing public and private donors. Are there outliers for either of these types of donors in 2000? What about in 2020? (Tip: box plots are great for distributions! Please treat year as factor and scale the nominal amount using "scale_y_log10(labels = scales::dollar)").

Answer

- Wrangling the data:

1) filter for the year 2000 or 2020

2) drop NAs

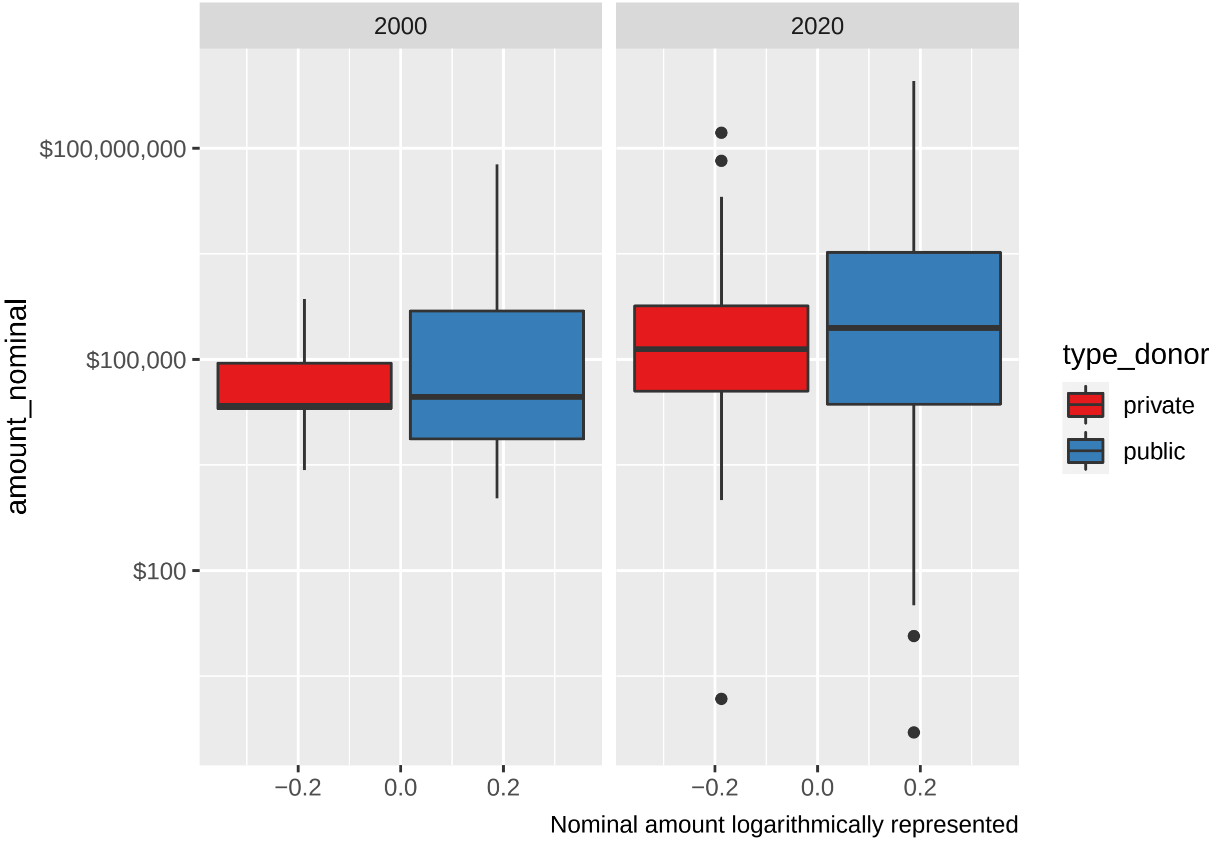
- Creating the box plot:

1) logarithmic scale used or else the boxplot becomes unreadable
2) scale fill brewer has palettes for people with color blindness

- Interpretation Outliers are present for both 2000 and 2020 and are visualized as individual points. Comparing 2000 with 2020, it seems that the distribution has become ‘wider’, i.e. observations are further away from each other

- Question: How do I remove the values on the x-as?

```
io_income_rs %>%
  filter(year == 2000 | year == 2020) %>%
  drop_na() %>%
  ggplot(mapping = aes(y=amount_nominal, x=0, fill=type_donor)) + geom_boxplot() + scale_y_log10(labels=
```



Question 4

Create four scatter plots, one for each issue area, containing all donation by year per donor type (Tip: use “facet_wrap(~issue_area)”). The shape of the points should reflect the donor type (Tip: use “geom_point(aes(shape = donor_type”))“). In each facet, please add a smoothed line to show the direction of the relationship (TIP: use”geom_smooth()“). The smoothed lines should be colored”red”.

Answer

- Wrangling the data:

1) drop NAs

- Creating the scatter (aka point) plot:

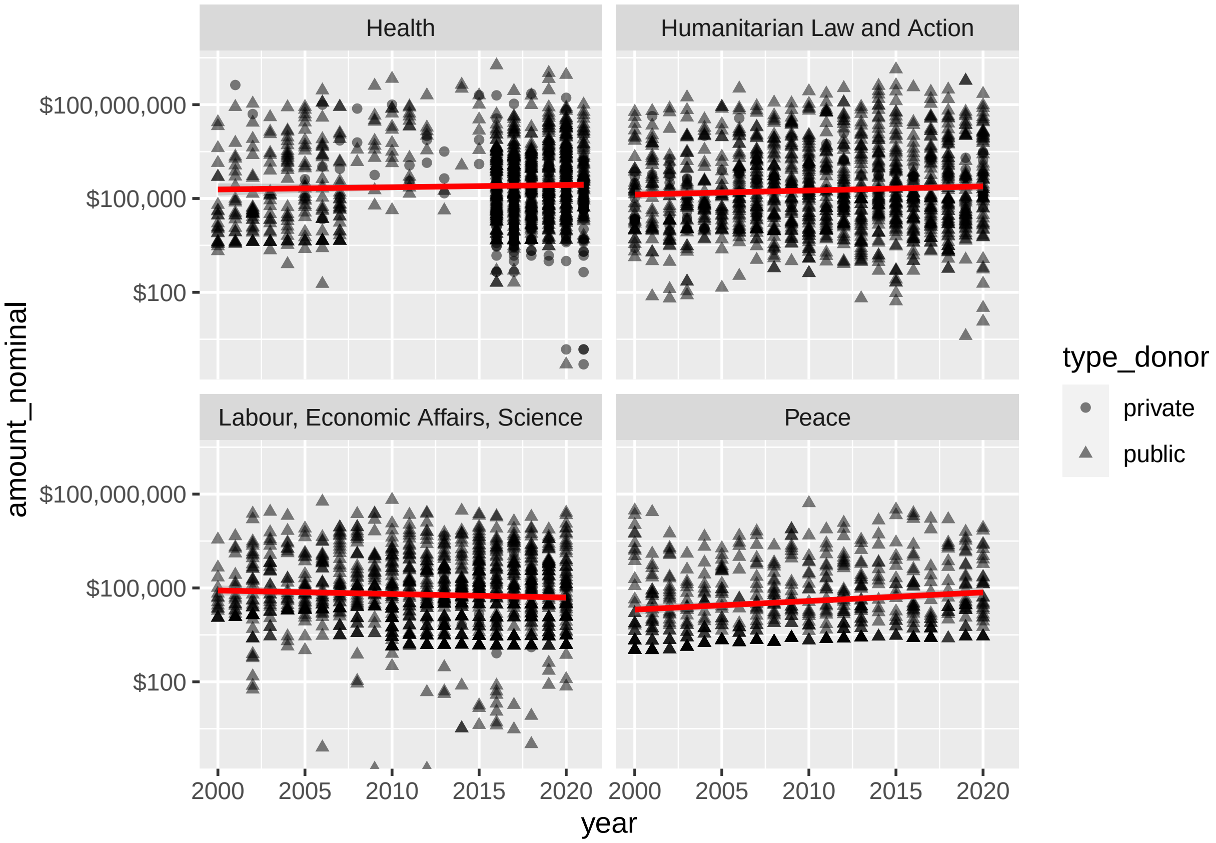
1) In order to prevent overplotting, we make points transparent through alpha = 0.5
2) Create a red line following a generalized linear model (glm)
3) Change scale of amount nominal to logarithmic scale

```
io_income_rs %>%
  drop_na() %>%
  ggplot(mapping = aes(x=year, y=amount_nominal)) + geom_point(aes(shape=type_donor),  alpha=0.5)+facet_
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
## Transformation introduced infinite values in continuous y-axis
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```


Health
Humanitarian Law and Action
Labour, Economic Affairs, Science
Peace
$100,000,000
$100,000
$100
$100,000,000
$100,000
$100
amount_nominal
year
type_donor
● private
▲ public