Data visualisation (computer session)

Daniele Rotolo

Introductory Data Science for Innovation (995N1) – Weeks 3, 11 October 2021

Objectives

- To familiarise with `ggplot` (we will use a dataset on firms' publishing activity and R&D expenditures)
- To explore a few network layout algorithms

Working with `ggplot`

We will rely on a sample of data from Camerani et al. (2018) (https://www.sussex.ac.uk/webteam/gateway/file.php?name=2018-21-swps-camerani-et-al.pdf&site=25). This sample includes 391 firms in the Pharmaceutical and Healthcare sector listed in the 2014 EU Industrial R&D Investment Scoreboard (https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/2014-eu-industrial-rd-investment-scoreboard). The dataset includes a range of variables:

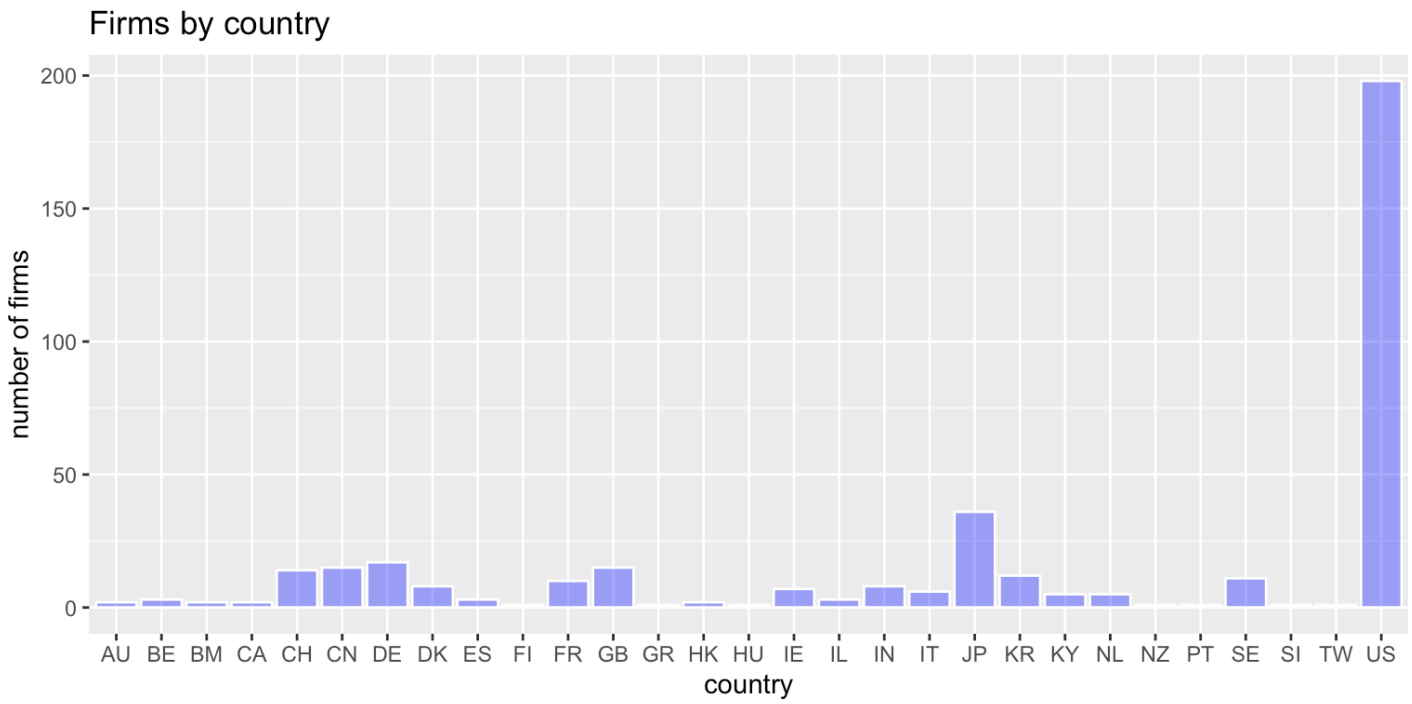
- `ID` : a firm's unique identifier
- `isocountrycode` : a firm's headquarter location (country-level)
- `rd2011` to `rd2015` : a firm's R&D expenditure from 2011 to 2015
- `ns2011` to `ns2015` : a firm's net sales from 2011 to 2015
- `emp2011` to `emp2015` : a firm's employees from 2011 to 2015
- `pub.2011` to `pubs.2015` : a firm's number of publications from 2011 to 2015

We first load the packages we need to visualise the data and we also load the data (please note the the working directory will be the directory where you save the ".Rmd" file)

```
rm(list=ls())
library(tidyverse)
library(GGally)
library(gghighlight)
library(patchwork)
my_data <- read_csv("scoreboard_firms_pharma_healthcare.csv")
```

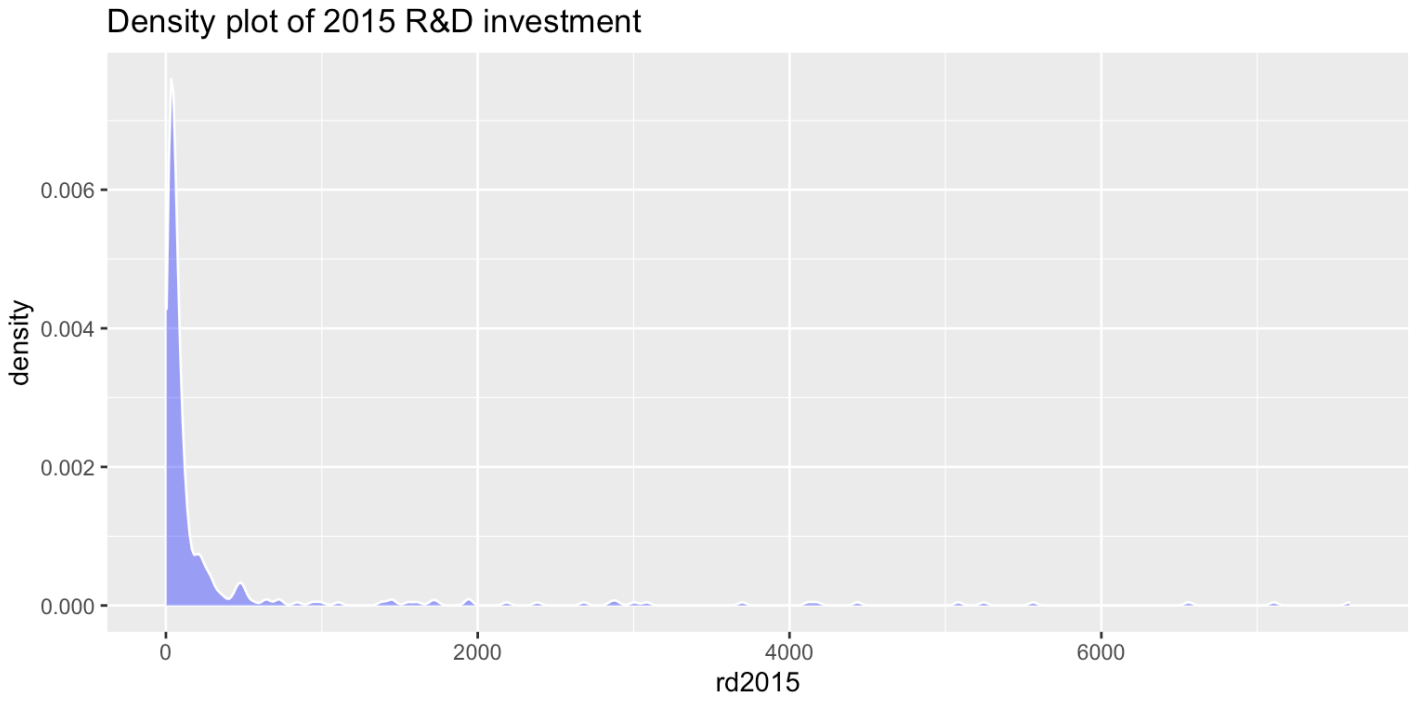
We start with examining the distribution of some variables in the dataset. In the case of `isocountrycode` , an histogram is more appropriate since this variable is categorical, while for all the remaining variables we can plot a density function.

```
ggplot(data = my_data, aes(isocountrycode)) +
  geom_histogram(stat = "count", color = "white",
                 fill = "blue", binwidth = 1, alpha = 0.4) +
  ggtitle("Firms by country") +
  xlab("country") + ylab("number of firms")
```



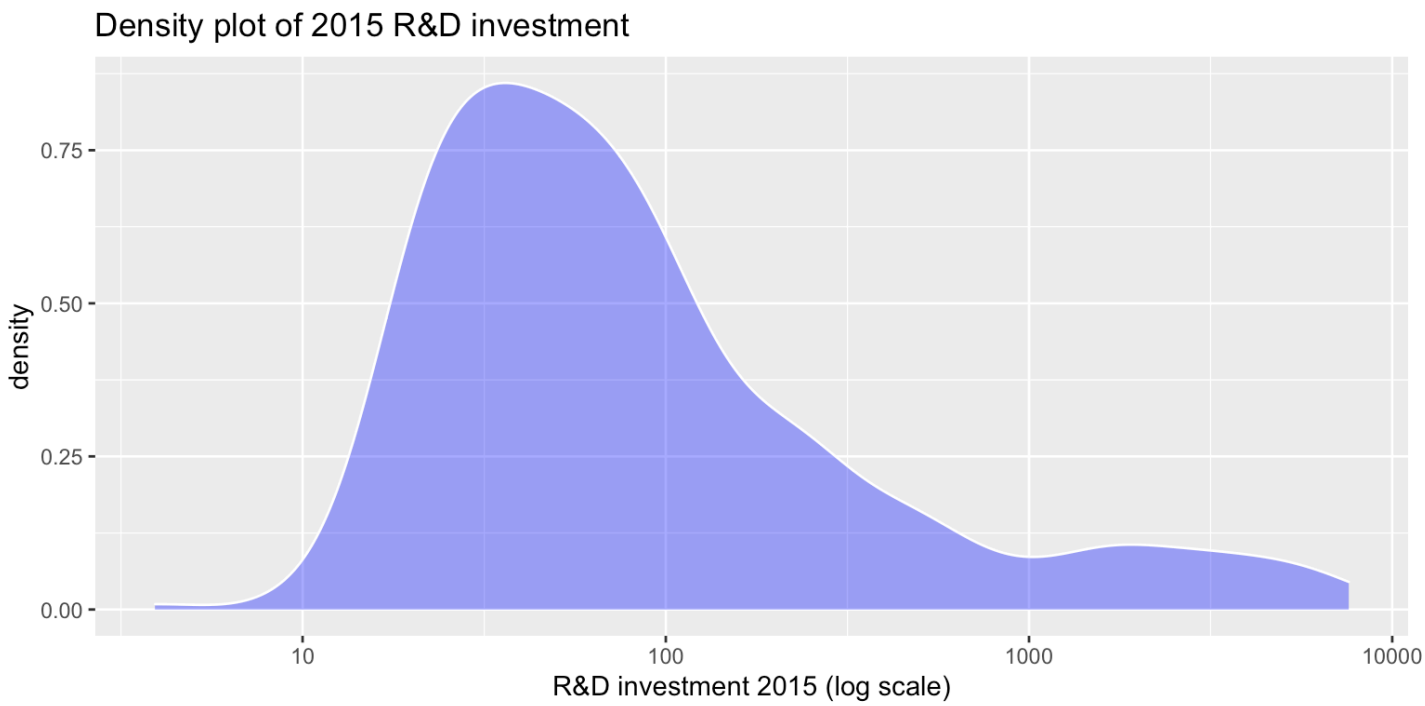
We now explore the remaining variables on R&D expenditure, net sales, employees, and publications in a given year. As an example, we select the year 2015.

```
ggplot(data = my_data, aes(rd2015)) +
  geom_density(color = "white", fill = "blue", alpha = 0.4) +
  ggtitle("Density plot of 2015 R&D investment")
```



The distribution is highly skewed. We ca transform the R&D investment variable using the log function.

```
ggplot(data = my_data, aes(rd2015)) +
  geom_density(color = "white", fill = "blue", alpha = 0.4) +
  scale_x_log10() +
  ggtitle("Density plot of 2015 R&D investment") +
  xlab("R&D investment 2015 (log scale)")
```



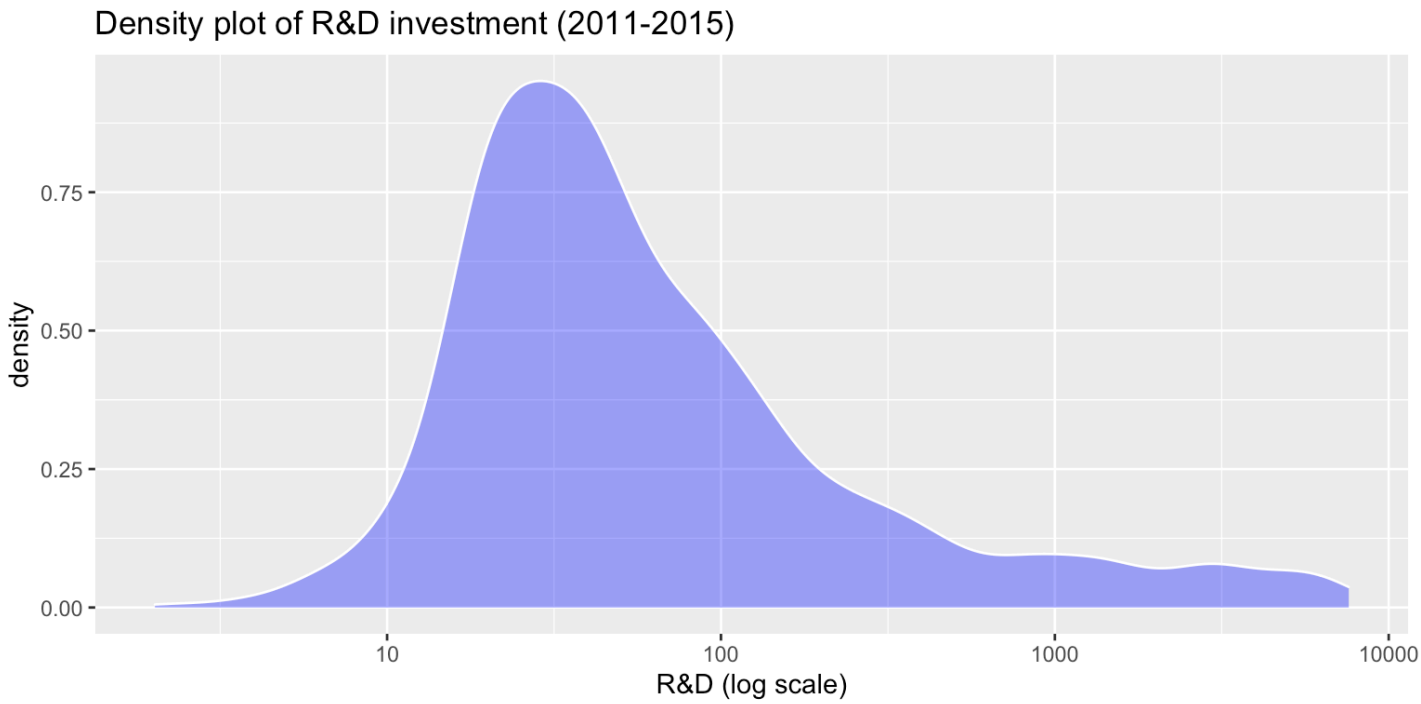
Exercise 1: Reproduce the density plot for the variable `pubs.2015` (5 minutes).

To exploit all year data, we need to transform our data into a tidy format. As an example, we focus on firms' R&D investment.

```
my_data_rd <- my_data %>%
  select(ID, rd2011, rd2012, rd2013, rd2014, rd2015) %>%
  pivot_longer(-ID, names_to = "year", values_to = "rd")

head(my_data_rd)
```

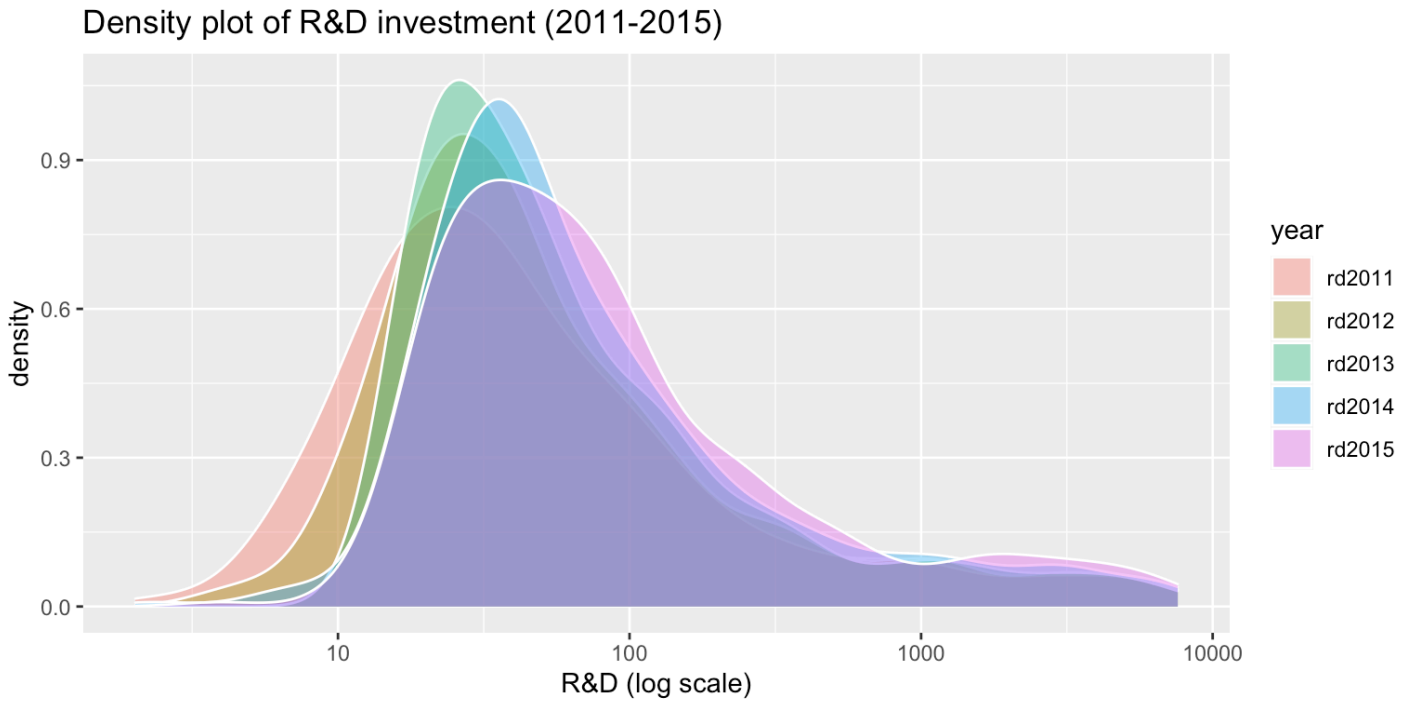
```
## # A tibble: 6 x 3
##   ID     year      rd
##   <chr>  <chr>  <dbl>
## 1 ID0001 rd2011 6657.
## 2 ID0001 rd2012 6737.
## 3 ID0001 rd2013 7174.
## 4 ID0001 rd2014 7234.
## 5 ID0001 rd2015 7106.
## 6 ID0002 rd2011 6566.
```

```
ggplot(data = my_data_rd, aes(rd)) +
  geom_density(color = "white", fill = "blue", alpha = 0.4) +
  scale_x_log10() +
  ggtitle("Density plot of R&D investment (2011-2015)") +
  xlab("R&D (log scale)")
```



The tidy structure allows us to explore our data by year and to generate automatically a legend in `ggplot2`.

```
ggplot(data = my_data_rd, aes(rd, fill = year)) +
  geom_density(color = "white", position = "identity", alpha = 0.4) +
  scale_x_log10() +
  ggtitle("Density plot of R&D investment (2011-2015)") +
  xlab("R&D (log scale)")
```



Exercise 2: Reproduce the density plot of the number of publications for each year (5 minutes).

We can now explore relationships between variables. To do so, we now need to transform the entire dataset into a tidy dataset.

```
head(my_data)
```

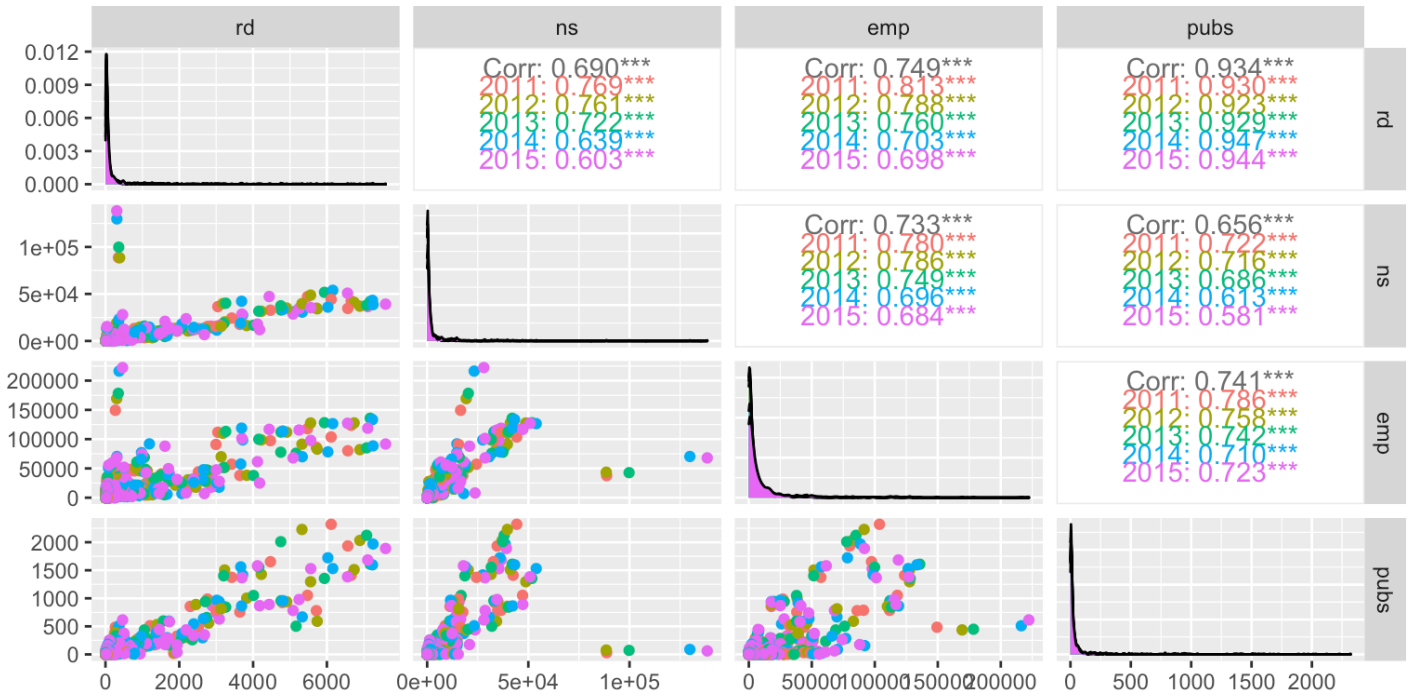
```
## # A tibble: 6 x 22
##   ID     isocountrycode rd2015 rd2014 rd2013 rd2012 rd2011 ns2015 ns2014 ns2013
##   <chr>  <chr>           <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1 ID0001 CH              7106.  7234.  7174.  6737.  6657. 36536. 43212. 41998.
## 2 ID0002 CH              7590.  7249.  7076.  6893.  6566. 39159. 38604. 38049.
## 3 ID0003 US              6559.  6159.  5934.  5558.  5473. 50811. 53898. 51709.
## 4 ID0004 US              5083.  5332.  5165.  5736.  5714. 28640. 30627. 31929.
## 5 ID0005 FR              5246   4812   4757   4909   4795  34542  33770  32951
## 6 ID0006 US              5562.  6026.  4750.  5324.  6115. 35422. 35969. 37404.
## # … with 12 more variables: ns2012 <dbl>, ns2011 <dbl>, emp2015 <dbl>,
## #   emp2014 <dbl>, emp2013 <dbl>, emp2012 <dbl>, emp2011 <dbl>,
## #   pubs.2011 <dbl>, pubs.2012 <dbl>, pubs.2013 <dbl>, pubs.2014 <dbl>,
## #   pubs.2015 <dbl>
```

```
my_data_rd <- my_data %>%
  select(ID, rd2011:rd2015) %>%
  pivot_longer(-ID, names_to = "year", values_to = "rd") %>%
  mutate(year = gsub("rd", "", year))

my_data_ns <- my_data %>%
  select(ID, ns2011:ns2015) %>%
  pivot_longer(-ID, names_to = "year", values_to = "ns") %>%
  mutate(year = gsub("ns", "", year))

my_data_emp <- my_data %>%
  select(ID, emp2011:emp2015) %>%
  pivot_longer(-ID, names_to = "year", values_to = "emp") %>%
  mutate(year = gsub("emp", "", year))

my_data_pub <- my_data %>%
  select(ID, pubs.2011:pubs.2015) %>%
  pivot_longer(-ID, names_to = "year", values_to = "pubs") %>%
  mutate(year = gsub("pubs.", "", year))

my_data_tidy <- my_data_rd %>%
  full_join(., my_data_ns, by = c("ID", "year")) %>%
  full_join(., my_data_emp, by = c("ID", "year")) %>%
  full_join(., my_data_pub, by = c("ID", "year")) %>%
  full_join(., my_data %>% select(ID, isocountrycode), by = c("ID"))

head(my_data_tidy)
```

```
## # A tibble: 6 x 7
##   ID      year     rd     ns    emp  pubs isocountrycode
##   <chr>   <chr> <dbl>  <dbl>  <dbl> <dbl> <chr>
## 1 ID0001 2011  6657. 42467. 123686  1411 CH
## 2 ID0001 2012  6737. 41094. 127724  1512 CH
## 3 ID0001 2013  7174. 41998. 135696  1608 CH
## 4 ID0001 2014  7234. 43212. 133413  1597 CH
## 5 ID0001 2015  7106. 36536. 118700  1684 CH
## 6 ID0002 2011  6566. 34593.  80129  1934 CH
```

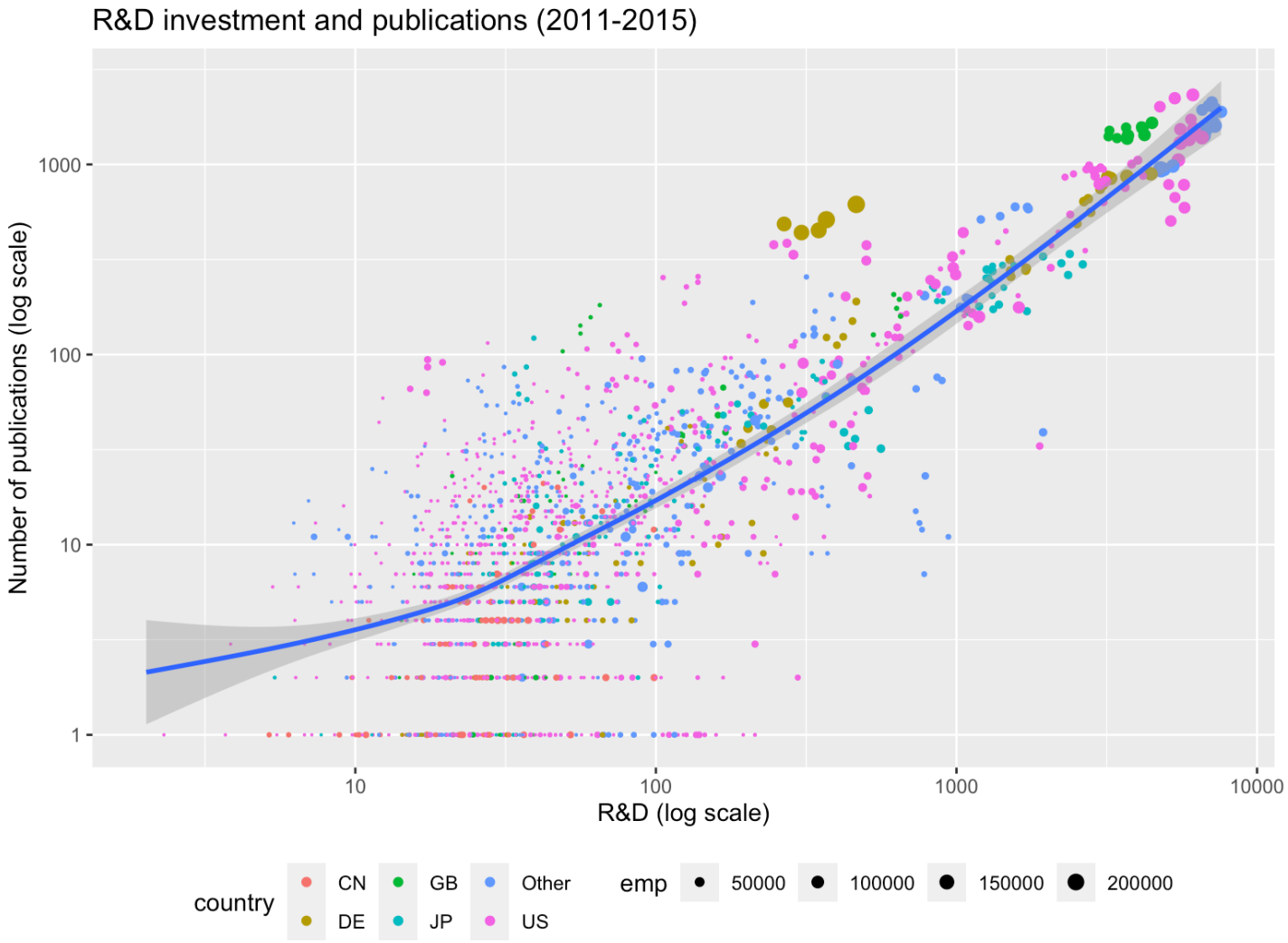
We can use the `GGally` package to explore relationships between variables by years using the new tidy data structure.

```
ggpairs(my_data_tidy, aes(color = year),
        columns = c("rd", "ns", "emp", "pubs"))
```



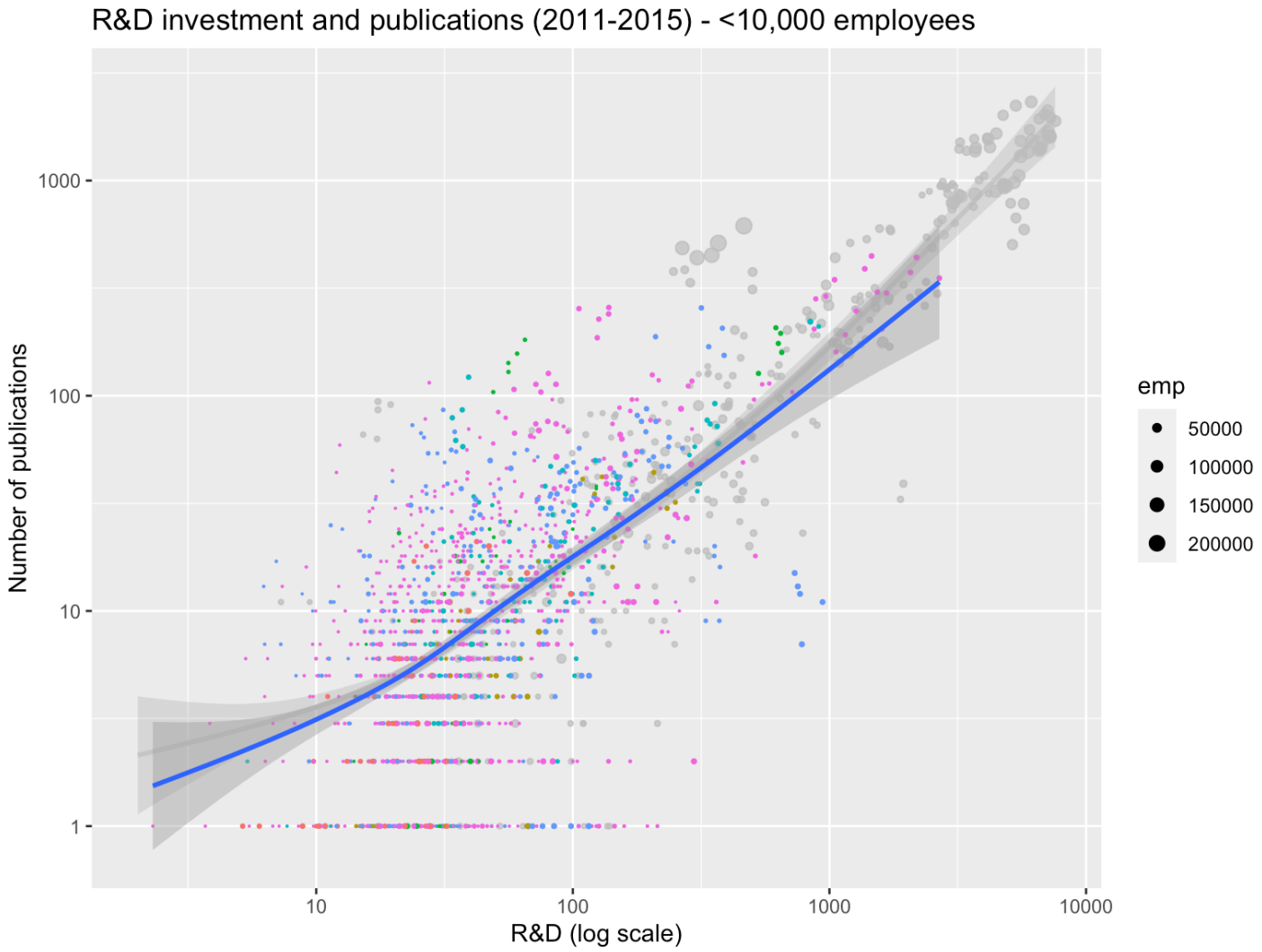
We can focus on the relationship between a firm's R&D investment and publications activity. We can also increase the size of the points on the basis of the number of employees and color them on the basis of country data. We will need to simplify the latter first.

```
my_data_tidy <- my_data_tidy %>%
  mutate(country = ifelse(isocountrycode != "US" &
                            isocountrycode != "CN" &
                            isocountrycode != "JP" &
                            isocountrycode != "DE" &
                            isocountrycode != "GB", "Other", isocountrycode))

ggplot(data = my_data_tidy, aes(x = rd, y = pubs+1)) +
  geom_point(aes(color = country,  size = emp)) +
  scale_size(range = c(0, 3)) +
  geom_smooth() +
  scale_x_log10() +
  scale_y_log10() +
  ggtitle("R&D investment and publications (2011-2015)") +
  xlab("R&D (log scale)") +
  ylab("Number of publications (log scale)") +
  theme(legend.position = "bottom")
```



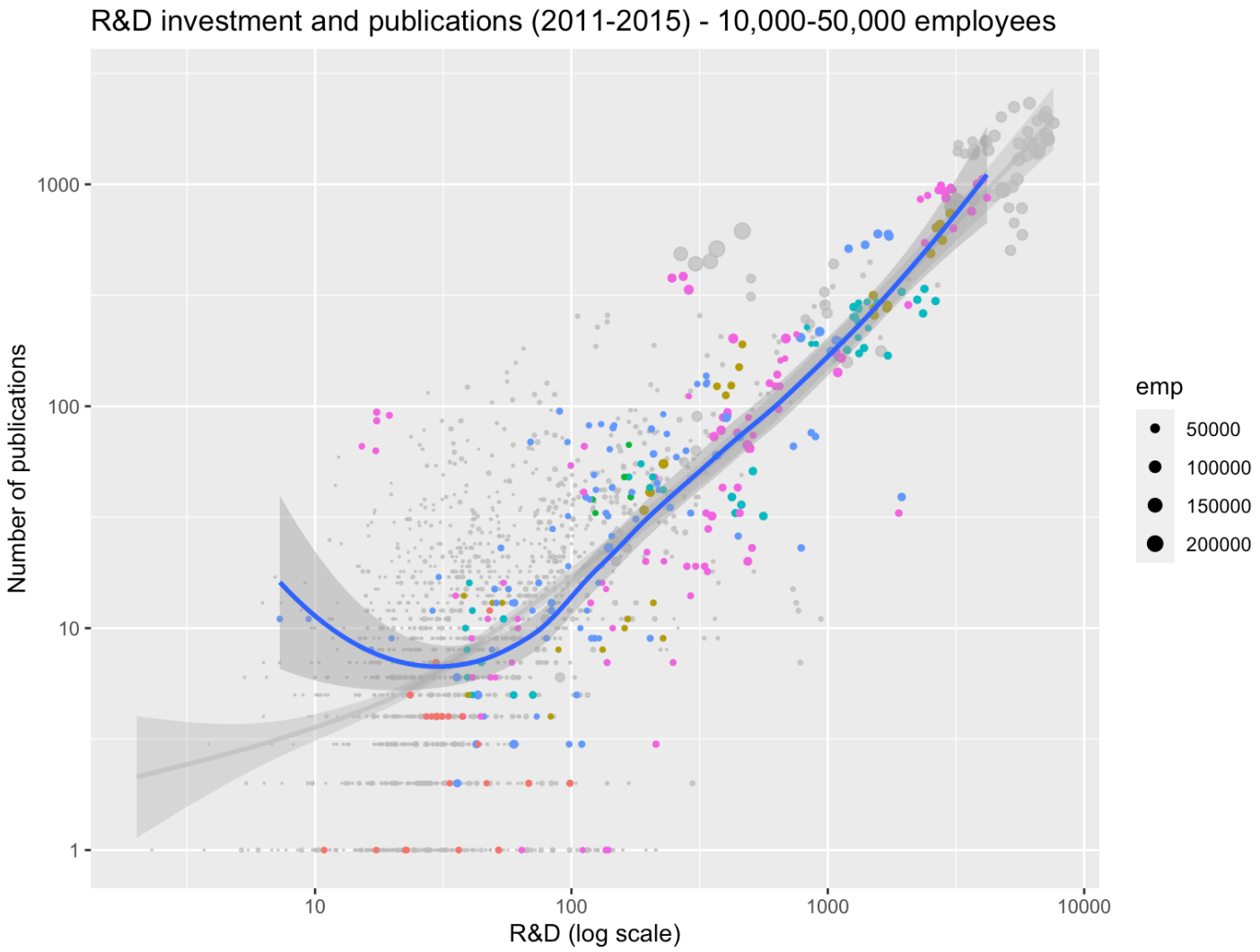
We can use the `gghighlight` package to identify firms with less than 10,000 employees…

```
ggplot(data = my_data_tidy, aes(x = rd, y = pubs+1)) +
  geom_point(aes(color = country,  size = emp)) +
  scale_size(range = c(0, 3)) +
  geom_smooth() +
  scale_x_log10() +
  scale_y_log10() +
  ggtitle("R&D investment and publications (2011-2015) - <10,000 employees") +
  xlab("R&D (log scale)") +
  ylab("Number of publications") +
  gghighlight(emp < 10000, keep_scales = T)
```



… or firms with 10,000-50,000 employees…

```
ggplot(data = my_data_tidy, aes(x = rd, y = pubs+1)) +
  geom_point(aes(color = country,  size = emp)) +
  scale_size(range = c(0, 3)) +
  geom_smooth() +
  scale_x_log10() +
  scale_y_log10() +
  ggtitle("R&D investment and publications (2011-2015) - 10,000-50,000 employees")
+
  xlab("R&D (log scale)") +
  ylab("Number of publications") +
  gghighlight(emp >= 10000 & emp <= 50000, keep_scales = T)
```



… or firms with more than 50,000 employees.

```
ggplot(data = my_data_tidy, aes(x = rd, y = pubs+1)) +
  geom_point(aes(color = country,  size = emp)) +
  scale_size(range = c(0, 3)) +
  geom_smooth() +
  scale_x_log10() +
  scale_y_log10() +
  ggtitle("R&D investment and publications (2011-2015) - >50,000 employees") +
  xlab("R&D (log scale)") +
  ylab("Number of publications") +
  gghighlight(emp > 50000, keep_scales = T)
```

We can combine all these charts using the `patchwork` package.

```
g1 <- ggplot(data = my_data_tidy, aes(x = rd, y = pubs+1)) +
  geom_point(aes(color = country,  size = emp)) +
  scale_size(range = c(0, 3)) +
  geom_smooth() +
  scale_x_log10() +
  scale_y_log10() +
  ggtitle("R&D investment and number of publications (2011-2015)") +
  xlab("R&D (log scale)") +
  ylab("Number of publications") +
  theme(legend.position = "bottom")

g2 <- g1 +
  theme(legend.position = "none",
        plot.title = element_text(size = 7)) +
  ggtitle("<10,000 employees") +
  gghighlight(emp < 10000, keep_scales = T)

g3 <- g1 +
  theme(legend.position = "none",
        plot.title = element_text(size = 7)) +
  ggtitle("10,000-50,000 employees") +
  gghighlight(emp >= 10000 & emp <= 50000, keep_scales = T)

g4 <- g1 +
  theme(legend.position = "none",
        plot.title = element_text(size = 7)) +
  ggtitle(">50,000 employees") +
  gghighlight(emp > 50000, keep_scales = T)

g1 / (g2 + g3 + g4) + plot_annotation(tag_levels = 'A')
```

The `face_wrap()` function is a very helpful tool to produce multiple charts on the basis of a categorical variable. We can produce a chart for each country - note we grouped countries into CN, DE, GB, JP, US, and Other.

```
ggplot(data = my_data_tidy, aes(x = rd, y = pubs+1)) +
  geom_point(aes(color = country,  size = emp)) +
  scale_size(range = c(0, 3)) +
  geom_smooth() +
  scale_x_log10() +
  scale_y_log10() +
  ggtitle("R&D investment and number of publications (2011-2015)") +
  xlab("R&D (log scale)") +
  ylab("Number of publications") +
  theme(legend.position = "bottom") +
  facet_wrap(~country)
```

Similarly, we can produce a char by year.

```
ggplot(data = my_data_tidy, aes(x = rd, y = pubs+1)) +
  geom_point(aes(color = country,  size = emp)) +
  scale_size(range = c(0, 3)) +
  geom_smooth() +
  scale_x_log10() +
  scale_y_log10() +
  ggtitle("R&D investment and number of publications (2011-2015)") +
  xlab("R&D (log scale)") +
  ylab("Number of publications") +
  theme(legend.position = "bottom") +
  facet_wrap(~year)
```

Exercise 3: Produce a chart that compares R&D investment and number of publications for UK firms (10 minutes).