# Phytoplankton - filter - group_by - summarize

### Introduction

In this exercise we show some basics of data handling in R tidyverse. E.g. to select rows in a dataset that fullfill some condition (e.g. all observations made in July and August), compute group means (e.g monthly or annual means), and make some simple plots.

Before you start, make sure that you have installed and loaded the package `tidyverse` package.

```
library(tidyverse)
```

### Read and subset the phytoplankton dataset

This dataset contains the biovolume and cyanobacteria densities for a number of lakes in Southern Sweden[1].

```
phyto <- read_csv("Phytoplankton.csv")
```

We want to present the observations of biovolume and cyanobacteria for summer values. Start by making a summary for the data to see what we got:

```
summary(phyto)
```

In the summary we see that there are observations that are made during other months than summer months. We need to filter out observations made in July or August:

```
Phyto_selected <- phyto %>%
  filter(Month %in% c(7, 8))
```

### Making group summaries

Next we check how many observations there are for each year and station. This can be done using `group_by` to define groups and `n()` to determine the number of observations:

```
Phyto_selected <- phyto %>%
  filter(Month %in% c(7, 8)) %>%
  group_by(Station_name,
           Year) %>%
  summarize(n = n())
```

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

To see the dataset use \texttt{View(Phyto_selected)}. For some of the lakes there is only one observation per year, while others often have two.

### Compute mean values per group

so, if we select all observations for July and August we get some years with two observations and others with only one. To proceed with plotting and analysis we often prefer to have the same amount of observations for each year. An option is, then, to compute mean values for the years where two observations are available, i.e. compute annual means:

[1]This data is part of the Swedish environmental monitoring of phytoplankton in lakes and accessible for everyone through SLU's open data, https://miljodata.slu.se/mvm/

```
Phyto_selected <- phyto %>%
  filter(Month %in% c(7, 8)) %>%
  group_by(Station_name,
           Year) %>%
  summarize(Biovolume = mean(Biovolume),
            `Cyanobacteria (mm3/l)` = mean(`Cyanobacteria (mm3/l)`),
            n = n())
```

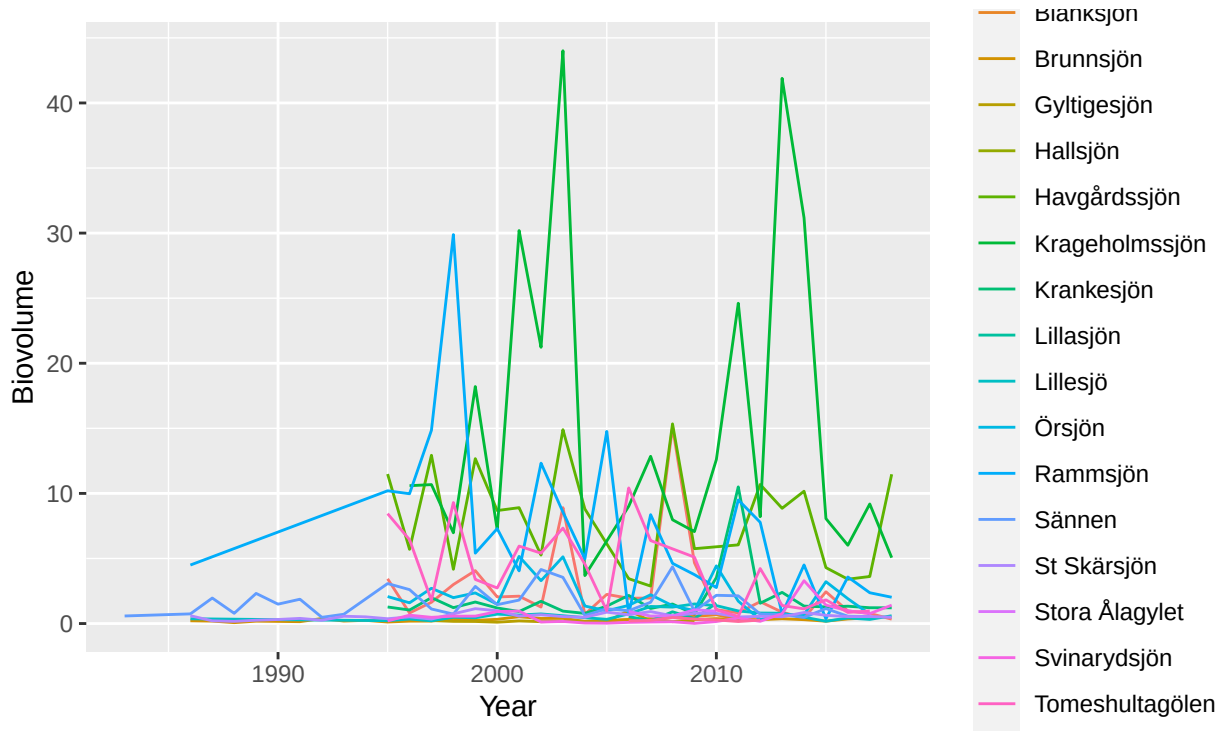
```
## `summarise()` has grouped output by 'Station_name'. You can override using the `.groups` argument.
```

Use again `View()` to see the results of your computations.

### Plot your data

To plot the annual data in one time series plots for all lakes we can use ggplot. Make sure to load the ggplot2 package first (which is loaded automatically when you load the tidyverse package). `x=` defines what is on the x-axis and `y=` the y-axis. `group` makes different lines for each station and texttt{color} sets different colors to each of the lines. The `geom_line` produces line plots.

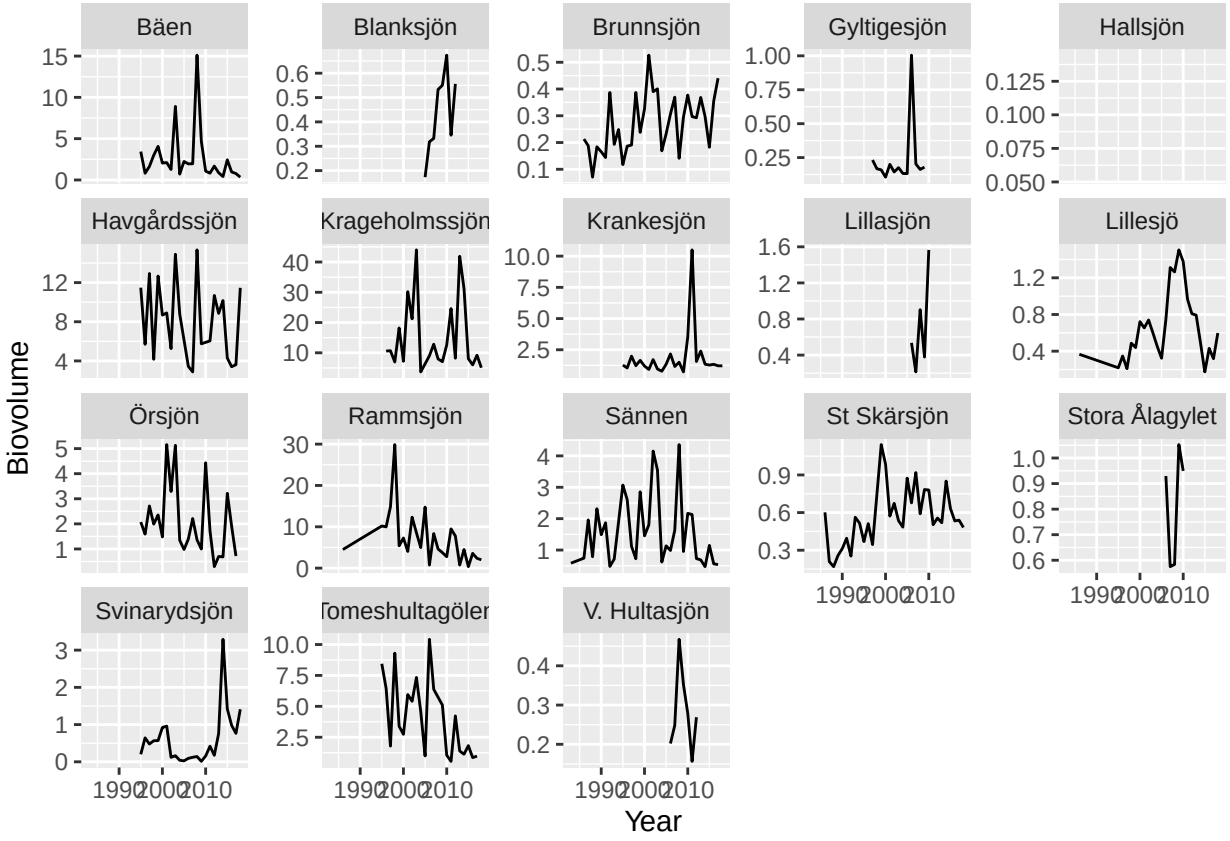
```
Phyto_selected %>%
  ggplot(aes(x = Year,
             y = Biovolume,
             group = Station_name,
             color = Station_name)) +
  geom_line()
```




**Task: Test what happens when you replace geom_line() with geom_point().**

Unfortunately it is difficult to see all lines and especially the series with low values are hard to identify. Maybe it is better to make one plot per series. We can obtain that by a `facet_wrap` by station name. The `scales="free_y"` let each plot have different scales on the y-axis.

```
Phyto_selected %>%
  ggplot(aes(x = Year,
             y = Biovolume)) +
  geom_line() +
  facet_wrap(~Station_name,
             scales = "free_y")
```




**Task: Test what happens when you remove the scales="free_y" statement**

**Exercises A**

1. Make a plot for Cyanobacteria in the same way as for Biovolume above.
2. In the plots created we see that many of the lakes only have short series and we decide that these should not be included. For this continue with the code you already have for `phyto_selected`, i.e. the data with annual means, and determine the number of observations per station for all years.
3. Filter the series in order to get a list with series (station names) that have more than 10 observations.

**Exercises B**

For these start with the dataset `phyto` (not `phyto_selected`):

4. Select all observations from years 2016 to 2018 and months July and August.
5. Compute the three-year-mean (i.e. for the selected years 2016 - 2018) for each station for both Biovolume and Cyanobacteria.
6. Find the lakes that have a three-year mean higher than 0.1 for Cyanobacteria.

**Solutions to exercises A**

1.

```
Phyto_selected %>%
  ggplot(aes(x = Year,
             y = `Cyanobacteria (mm3/l)`)) +
  geom_line() +
  facet_wrap(~Station_name,
             scales = "free_y")
```

2. Count annual observations per station:

```
phyto %>%
  filter(Month %in% c(7, 8)) %>%
  group_by(Station_name,
           Year) %>%
  summarize(Biovolume = mean(Biovolume),
            `Cyanobacteria (mm3/l)` = mean(`Cyanobacteria (mm3/l)`),
            n = n()) %>%
  group_by(Station_name) %>%
  summarize(n = n())
```

3. Save the names of all long series in the tibble `long_series`:

```
long_series <- phyto %>%
  filter(Month %in% c(7, 8)) %>%
  group_by(Station_name,
           Year) %>%
  summarize(Biovolume = mean(Biovolume),
            `Cyanobacteria (mm3/l)` = mean(`Cyanobacteria (mm3/l)`),
            n = n()) %>%
  group_by(Station_name) %>%
  summarize(n = n()) %>%
  filter(n >= 10)
```

*Note: You can also decrease the dataset to series with more than 10 observations. The difference is to not do summarize first but use n() directly in filter:*

```
long_series <- phyto %>%
  filter(Month %in% c(7, 8)) %>%
  group_by(Station_name,
           Year) %>%
  summarize(Biovolume = mean(Biovolume),
            `Cyanobacteria (mm3/l)` = mean(`Cyanobacteria (mm3/l)`),
            n = n()) %>%
  group_by(Station_name) %>%
  filter(n() >= 10)
```

**Solutions to exerices B**

4.

```
phyto_new <- phyto %>%
  filter(Month %in% c(7, 8) & Year %in% c(2016, 2017, 2018))
```

5.

```
phyto_new <- phyto %>%
  filter(Month %in% c(7, 8) & Year %in% c(2016, 2017, 2018)) %>%
  group_by(Station_name) %>%
  summarize(Biovolume = mean(Biovolume),
            Cyanobacteria = mean(`Cyanobacteria (mm3/l)`))
```

6.

```
phyto_new <- phyto %>%
  filter(Month %in% c(7, 8) & Year %in% c(2016, 2017, 2018)) %>%
  group_by(Station_name) %>%
  summarize(Biovolume = mean(Biovolume),
            Cyanobacteria = mean(`Cyanobacteria (mm3/l)`)) %>%
  filter(Cyanobacteria > 0.1)
```