ggplot2 Introduction

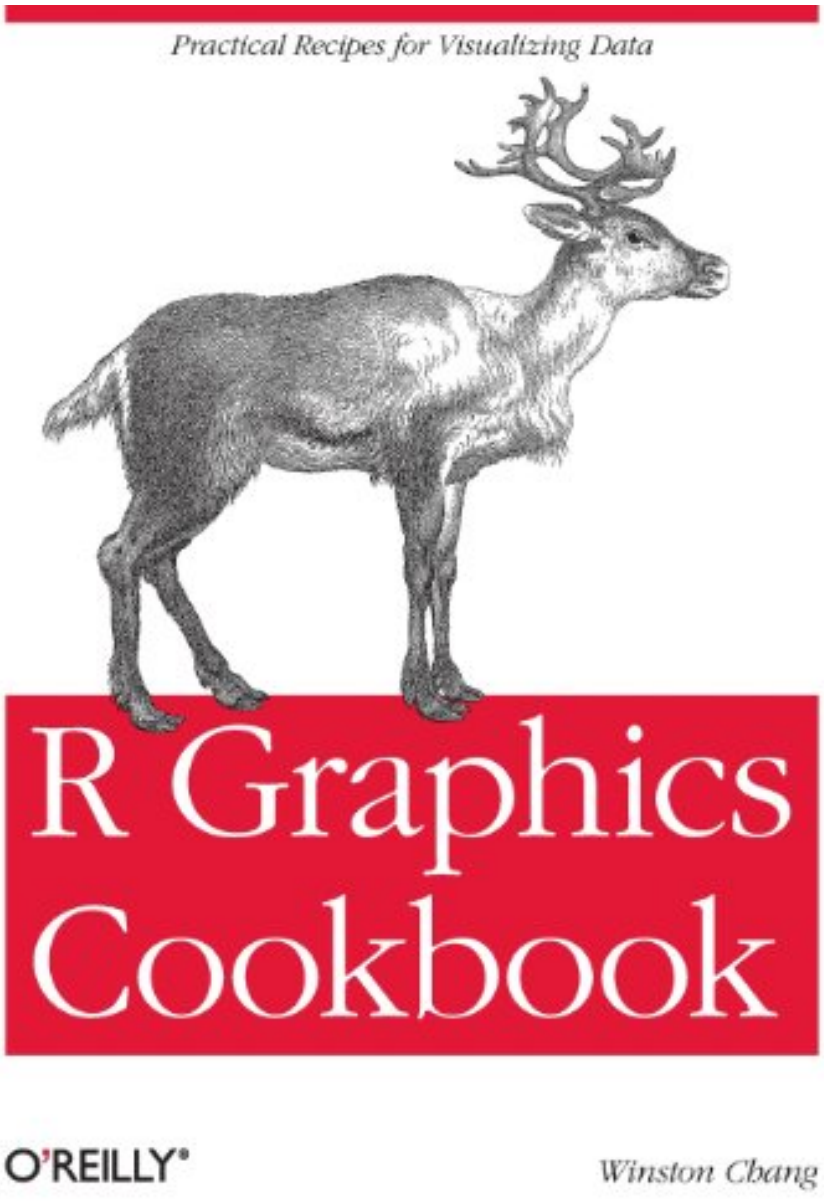
Jean-Baptiste Lecomte

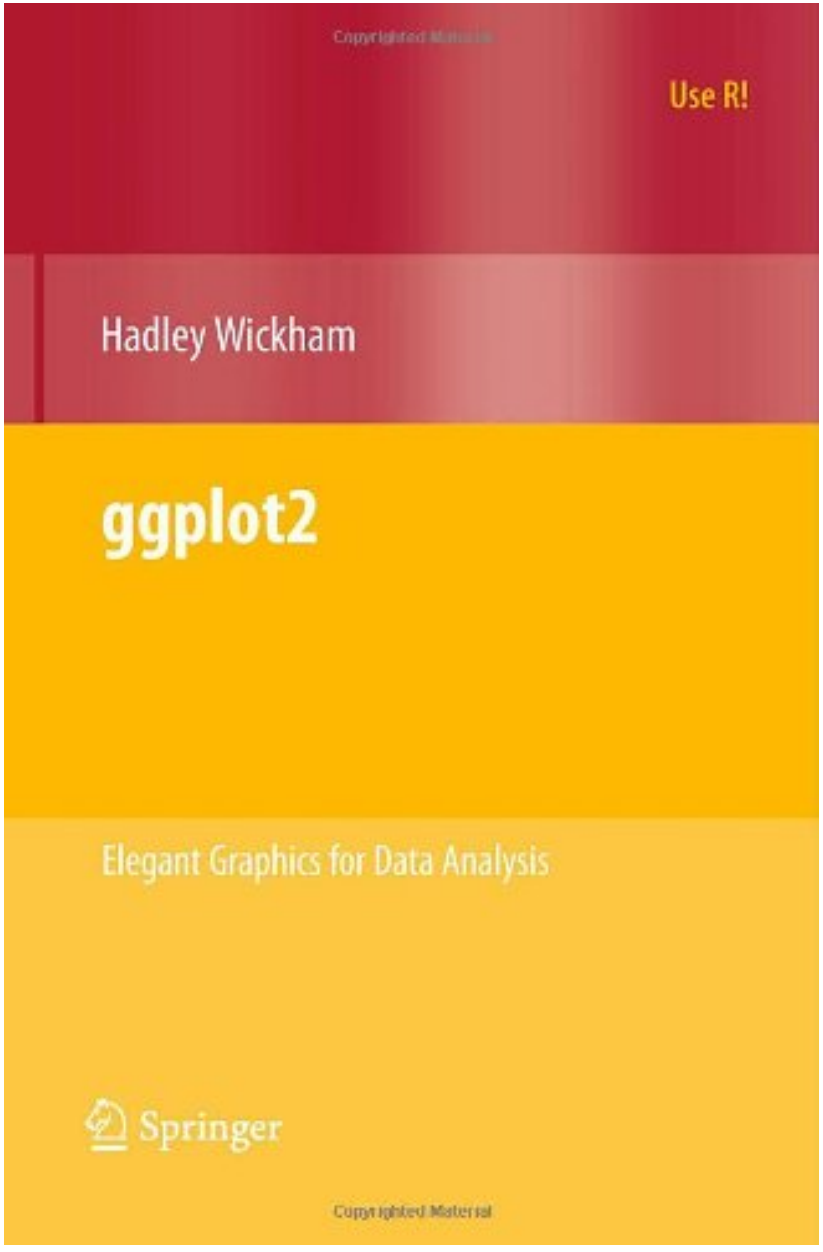
January 26, 2016

Introduction

- developped by Hadley Wickham (Rice University, Houston, USA)
- highly recommanded R packages to work with ggplot2: reshape and plyr (also developped by H. Wickham)
- first version called in 2007

Useful books





Online ressources

- ggplot2 official documentation:
  http://docs.ggplot2.org/current/
- R code related to ggplot2 cookbook:
  http://www.cookbook-r.com/Graphs/
- R code related to useR! ggplot2 book:
  http://ggplot2.org/book/
- Google groups to ask questions:
  ggplot2@googlegroups.com
- Github repository:
  https://github.com/yhat/ggplot/

Introduction

- based on new aesthetic principles
- based on *The grammar of graphics* developed by Wilkinson in 2005
- efficient way to produce simple graphics with a length reduction of R code

Forget about R base graphics:

`plot()`, `hist()`, `par()`, `layout()`, `points()`, `lines()`,`legend()`

Principle

ggplot2 is based on a **layer** system which can be used as objects.

Main layers

- data $\rightarrow$ raw data
- mapping $\rightarrow$ graphic projection
- geom $\rightarrow$ geometric objects (points, lines, polygons, ...)
- stat $\rightarrow$ statistics transformation (histogram, model)
- scale $\rightarrow$ aesthetics customization (color, shape, size, axes, legend)
- coord $\rightarrow$ coordinate system (axes, grid)
- facet $\rightarrow$ subdivision (lattice, trellis)

Base functions

ggplot2 is based on two functions:

1. `qplot()` for **q**uick **plot**
   - easy and fast, but too simple in most cases
   - `qplot(x, y, data=data)`

2. `ggplot()`
   - more complex but more powerful and flexible by adding `layers`
   - `ggplot(data=data, aes(x, y)) + layers`

Getting Started

Data format

Always work with a `data.frame`

Our data frame is based on the surveys XXXX and simulated data. Github repository:
https://github.com/JBLecomte/ggplot2-Introduction.git

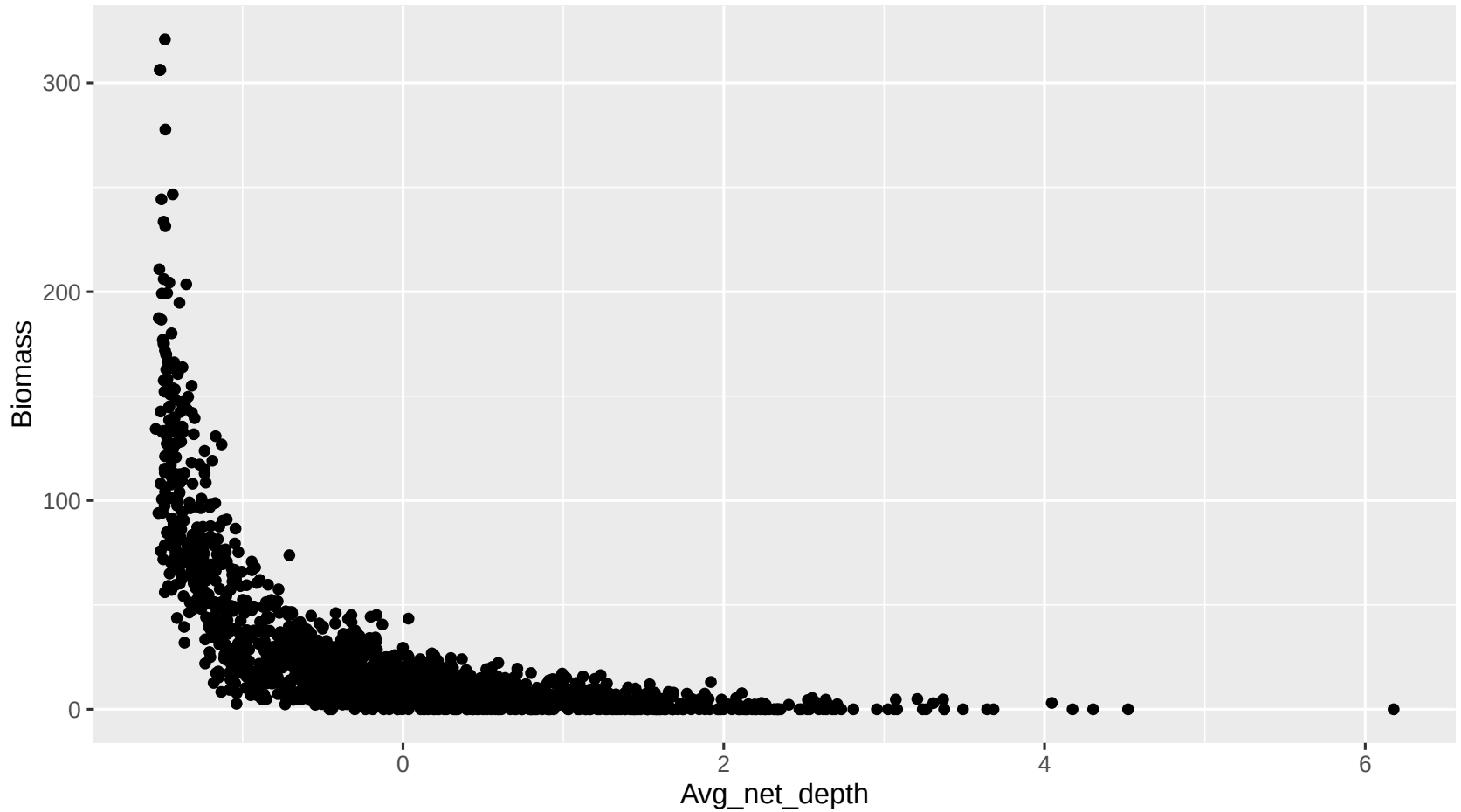
Getting Started

```
  str(df_data)
```

```
## 'data.frame': 1909 obs. of  18 variables:
##  $ Year            : int  2005 2005 2005 2005 2005 2005 2005 2005 2005 2005 ...
##  $ Month           : int  7 7 7 7 7 7 7 7 7 7 ...
##  $ DURATION_MINUTES: int  21 20 21 21 20 20 20 21 21 20 ...
##  $ AREA            : Factor w/ 2 levels "5AB","5CD": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Avg_net_depth   : num  -0.316 -0.435 -0.442 -0.234 -0.171 ...
##  $ Avg_net_temp    : num  0.3939 0.4339 0.3004 0.1335 -0.0267 ...
##  $ Date            : Date, format: "2005-07-06" "2005-07-06" ...
##  $ Lon             : num  -128 -128 -128 -128 -128 ...
##  $ Lat             : num  51.2 51.1 51.6 51.6 51.7 ...
##  $ X               : num  572025 570307 553665 551917 546338 ...
##  $ Y               : num  5668122 5665874 5717947 5719597 5723992 ...
##  $ X_km            : num  572 570 554 552 546 ...
##  $ Y_km            : num  5668 5666 5718 5720 5724 ...
##  $ Pres            : num  1 1 1 1 1 1 1 0 0 1 ...
##  $ Year_fac        : Factor w/ 5 levels "2005","2007",..: 1 1 1 1 1 1 1 1 1 1 ..
##  $ AREA_num        : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ nFish           : int  3 4 1 0 3 2 1 2 1 4 ...
##  $ Biomass         : num  7.17 9.77 2.15 0 7.28 ...
```

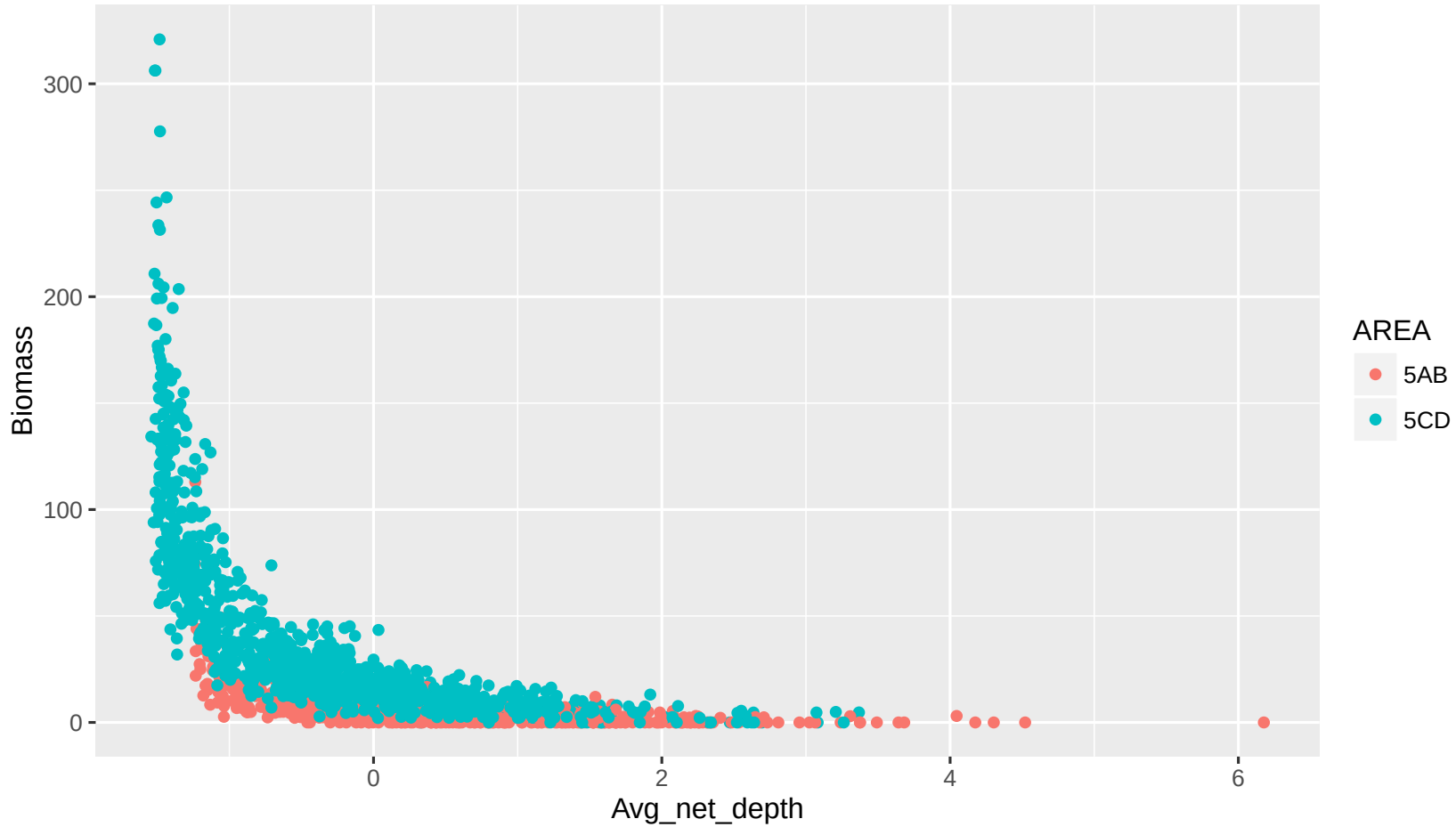
Scatter plot: Depth and Biomass

```
scatter_plot <- ggplot(data=df_data, aes(x=Avg_net_depth, y=Biomass)) +
                geom_point()
print(scatter_plot)
```



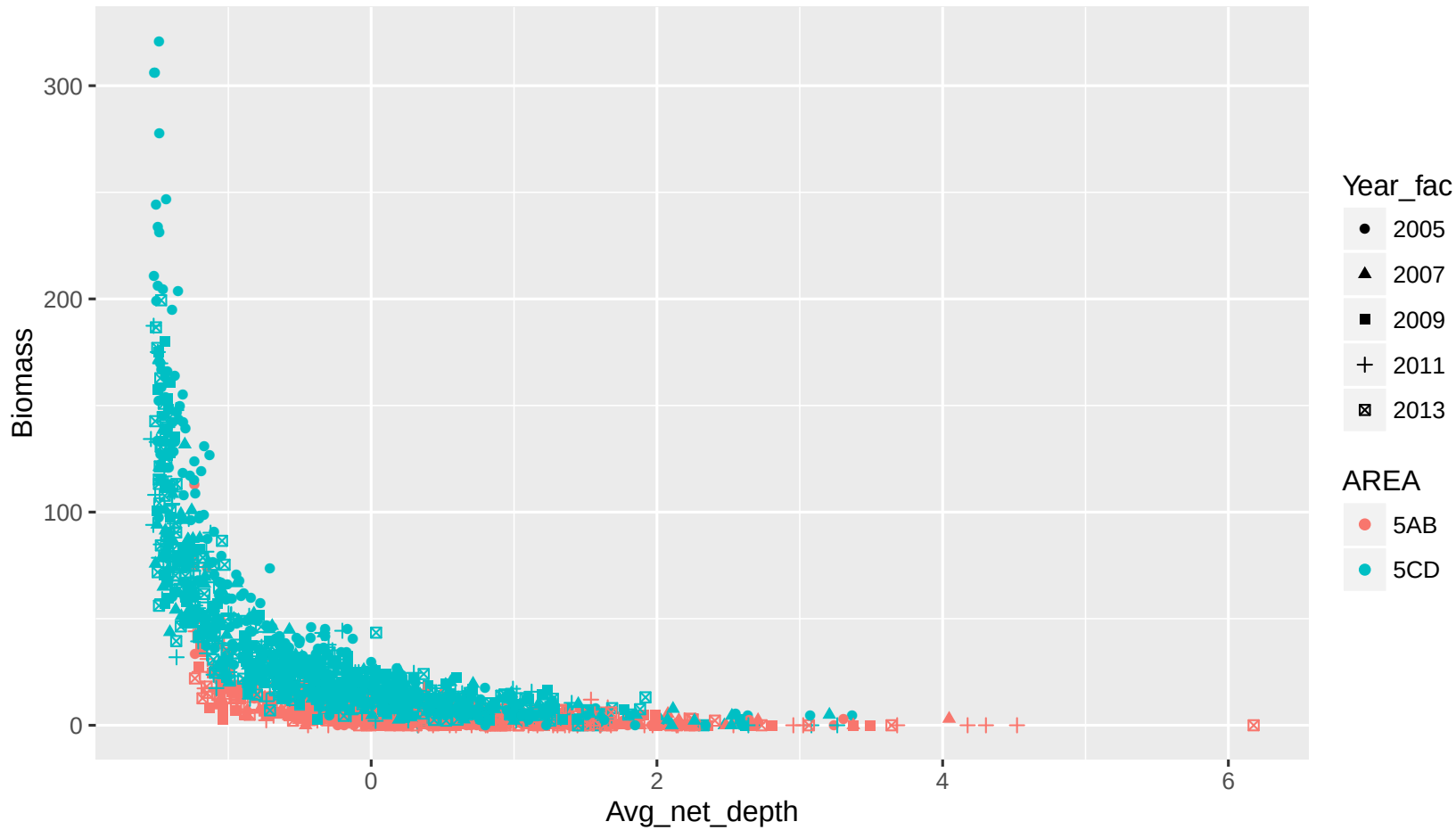
Scatter plot with color: Depth and Biomass

```
scatter_plot_color <- ggplot(df_data, aes(x=Avg_net_depth, y=Biomass,
                                          color=AREA)) +
geom_point()
print(scatter_plot_color)
```



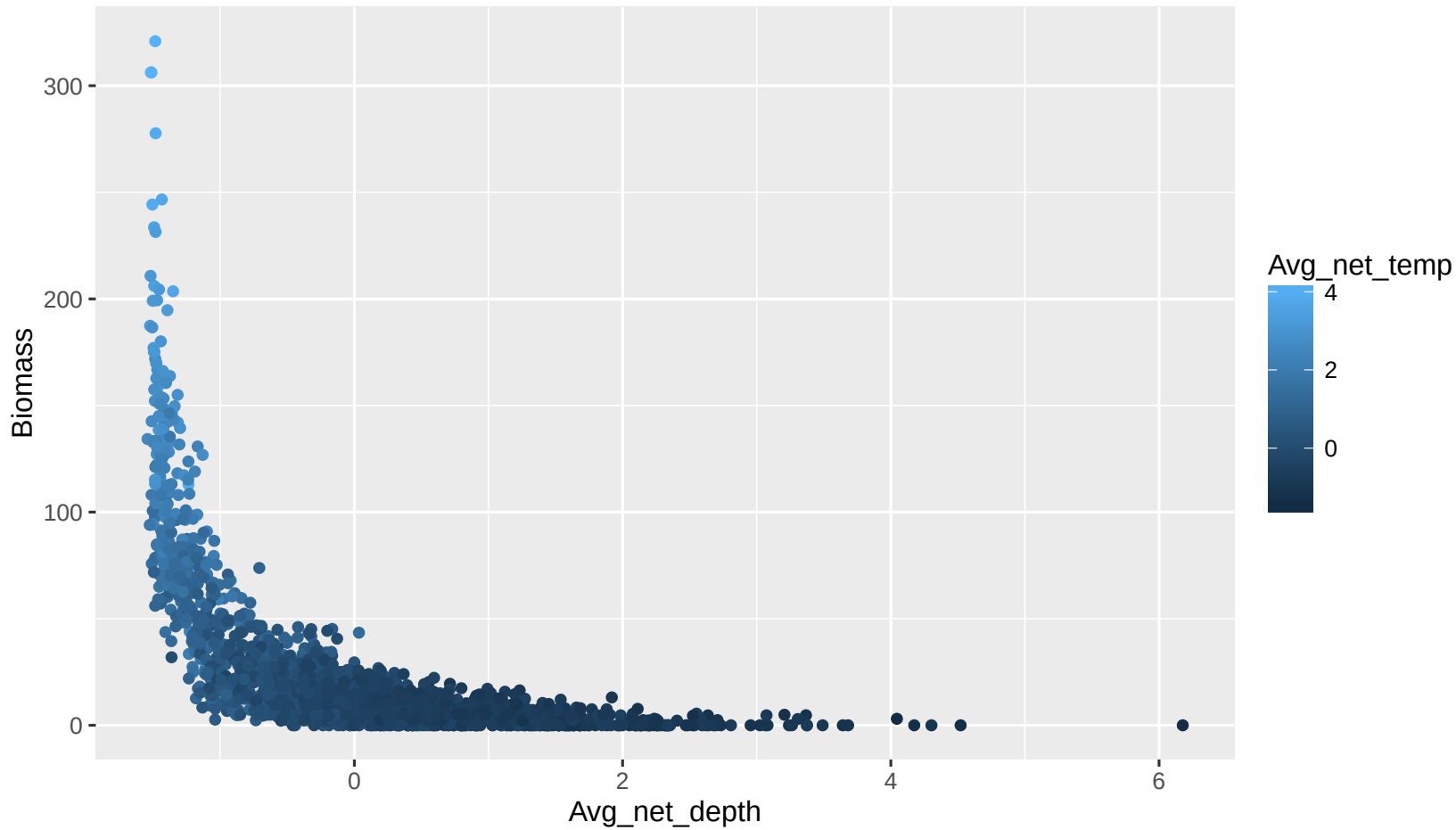
Scatter plot with shape: Depth and Biomass

```
scatter_plot_shape <- ggplot(df_data, aes(x=Avg_net_depth, y=Biomass,
                                          color=AREA, shape=Year_fac)) +
geom_point()
print(scatter_plot_shape)
```



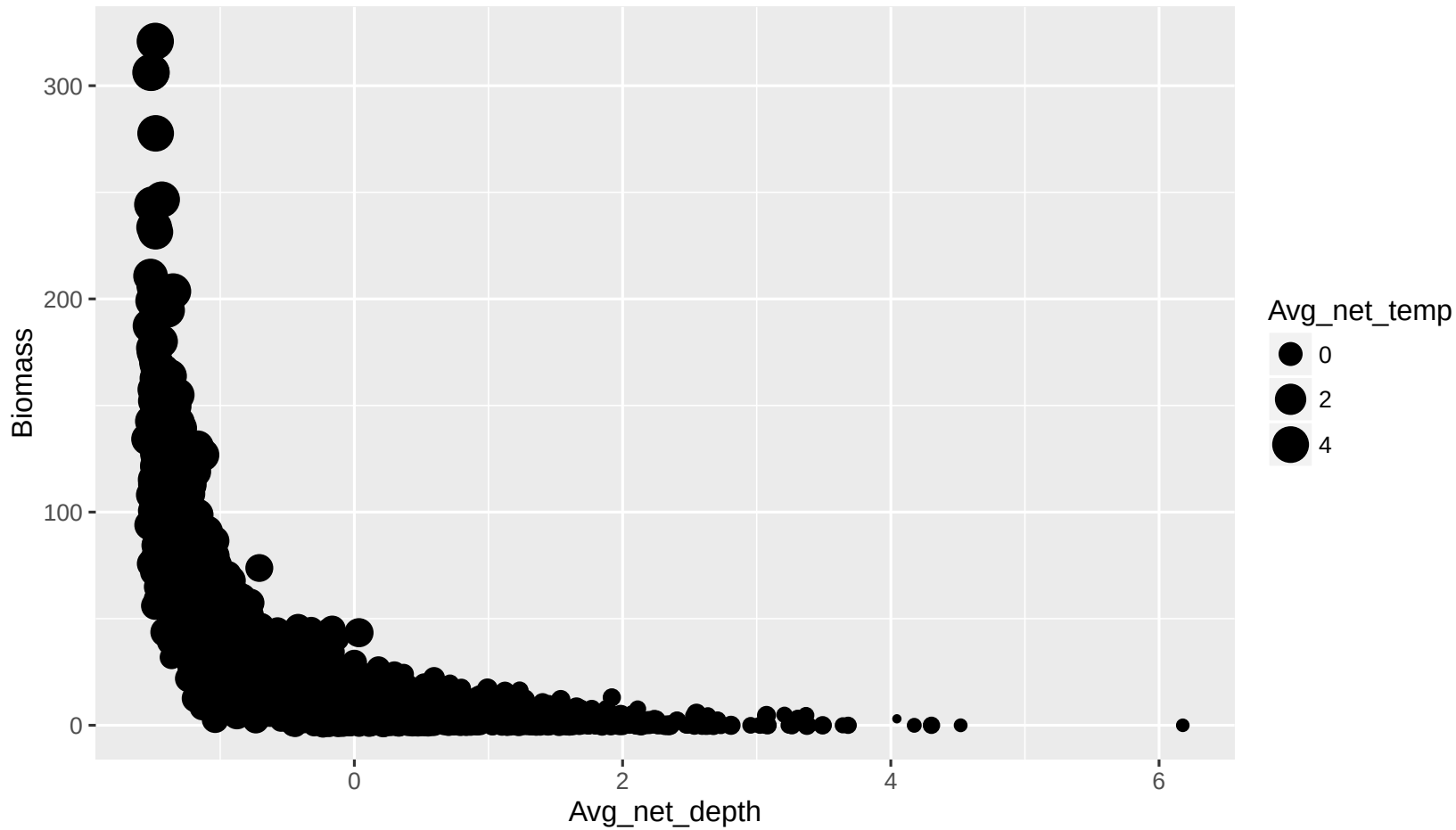
Scatter plot with continuous color: Depth and Biomass

```
scatter_plot_color_cont <- ggplot(df_data, aes(x=Avg_net_depth, y=Biomass
                                               color=Avg_net_temp)) +
geom_point()
print(scatter_plot_color_cont)
```



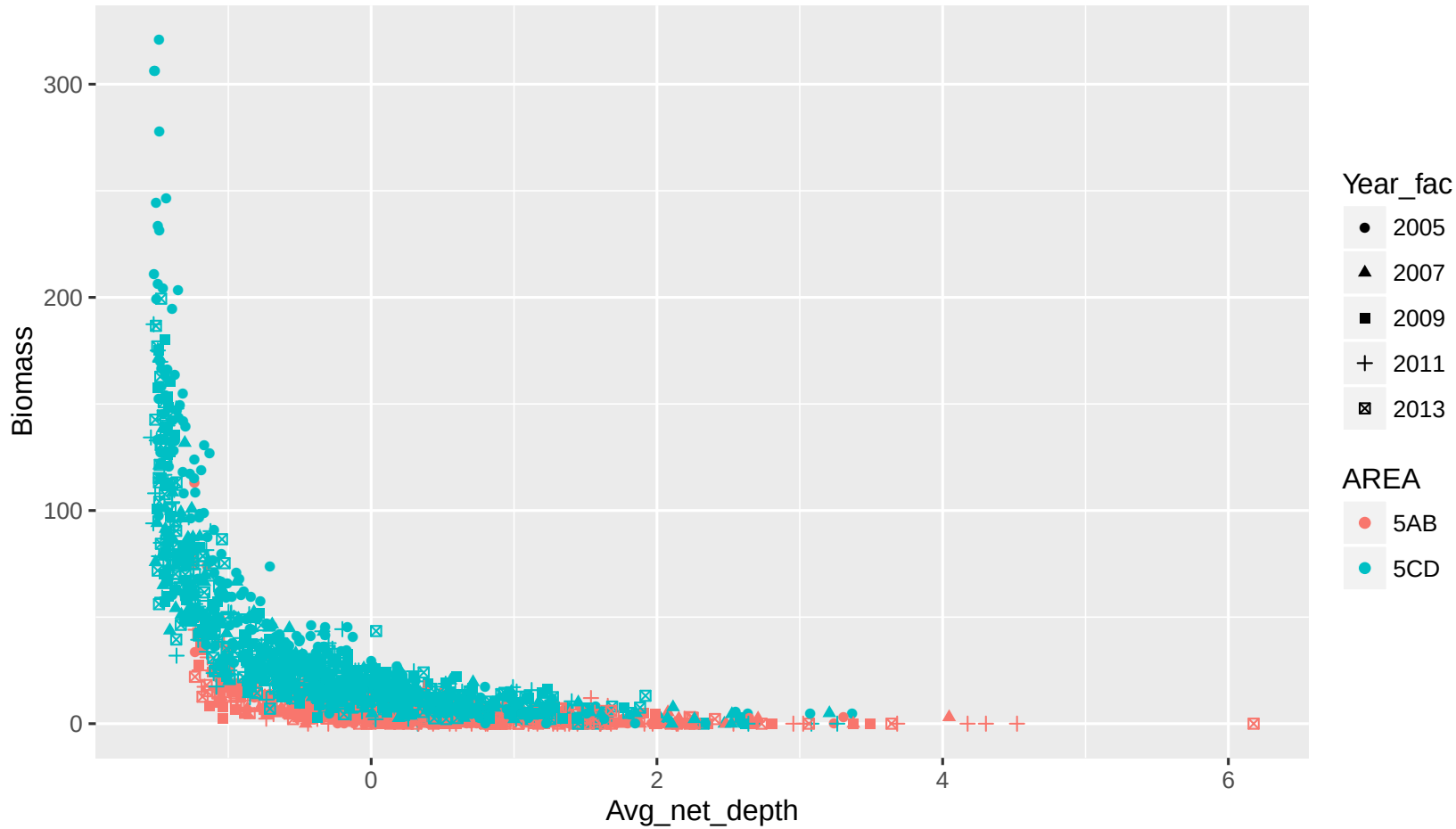
Scatter plot with size: Depth and Biomass

```
scatter_plot_area <- ggplot(df_data, aes(x=Avg_net_depth, y=Biomass,
                                         size=Avg_net_temp)) +
geom_point()
print(scatter_plot_area)
```



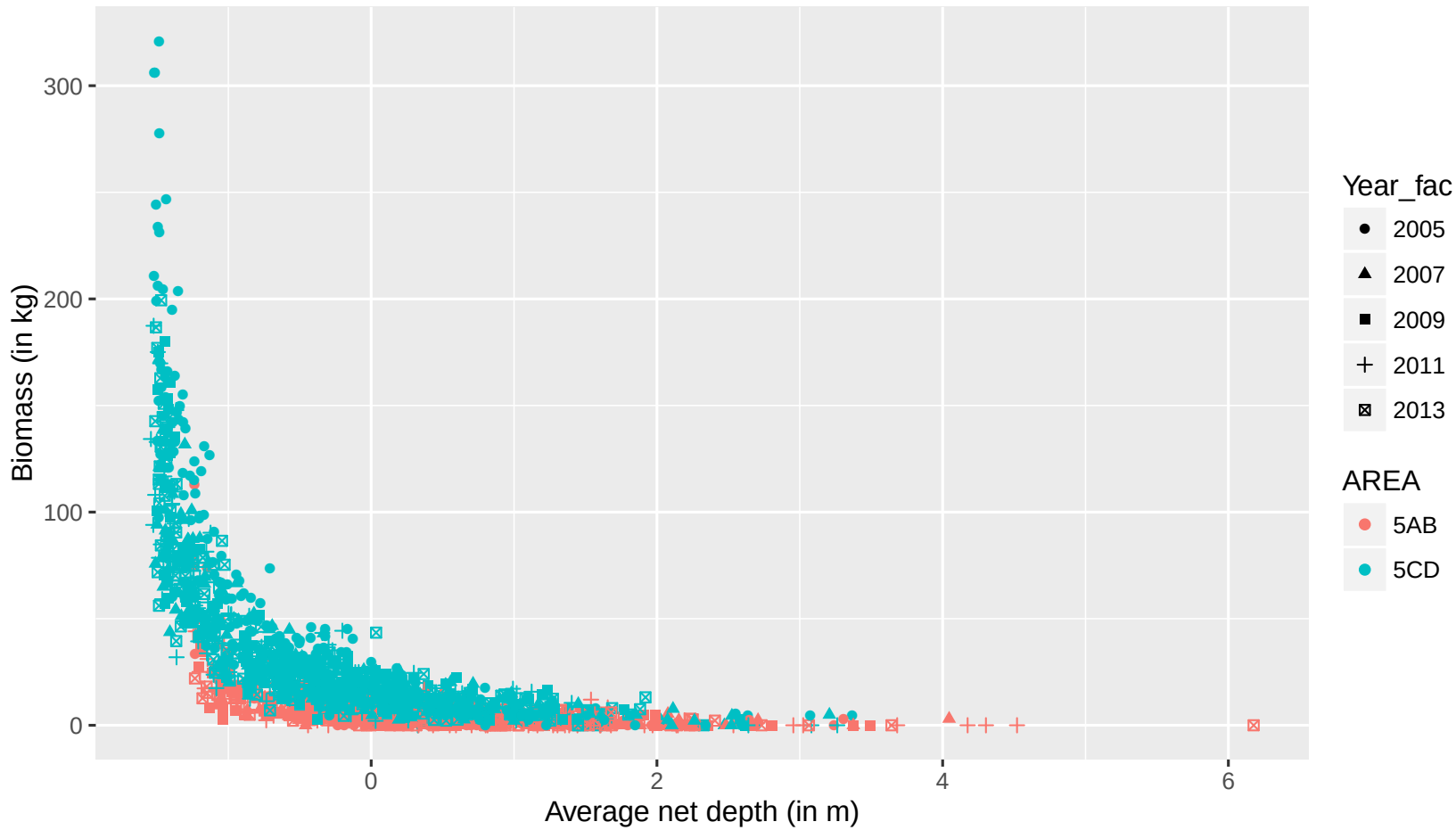
Improvement of a plot

```
print(scatter_plot_shape)
```



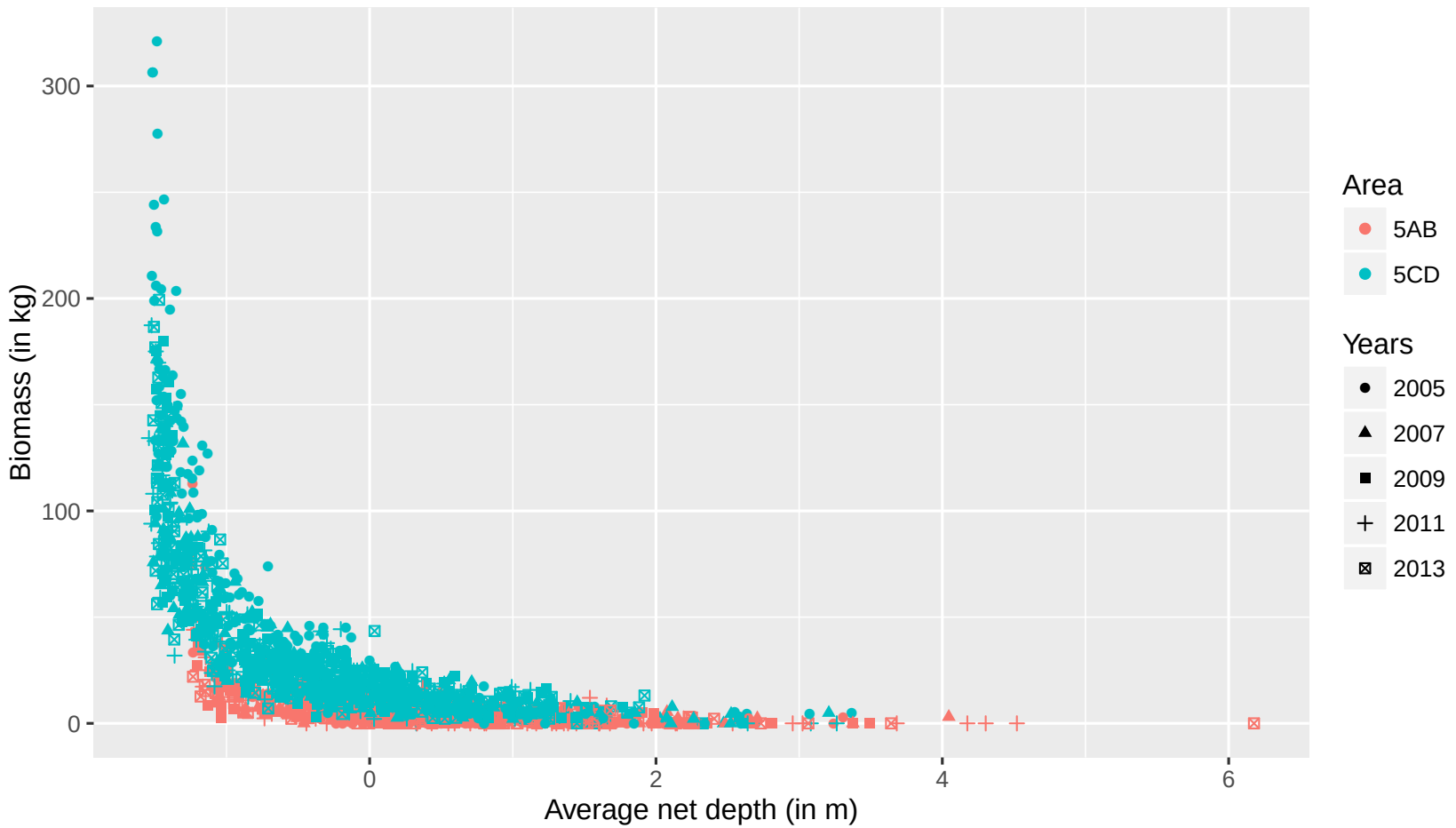
Improvement of a plot: axes names

```
scatter_plot_shape_imp1 <- scatter_plot_shape +
xlab('Average net depth (in m)') + ylab('Biomass (in kg)')

print(scatter_plot_shape_imp1)
```



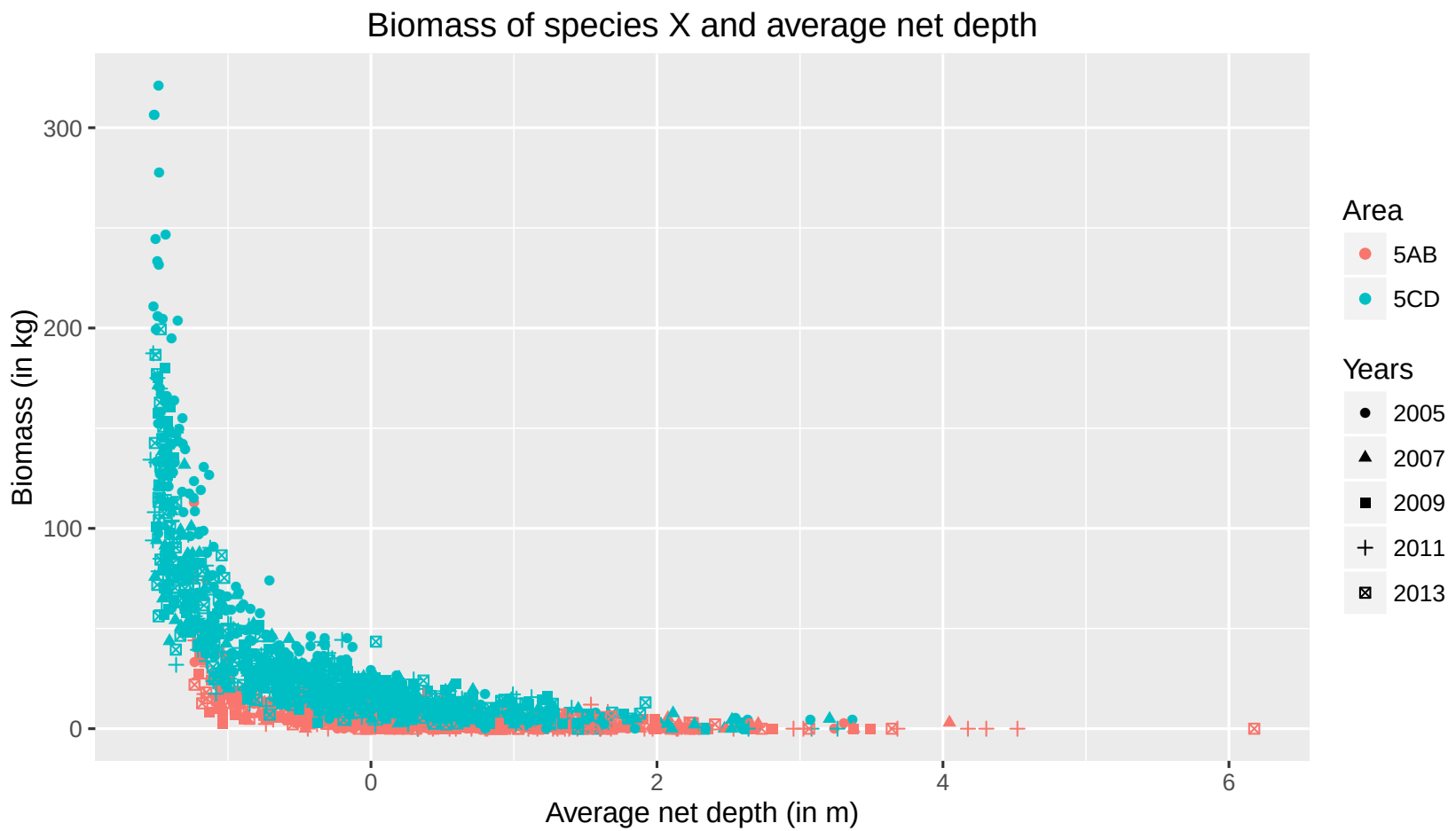
Improvement of a plot: legend titles

```
scatter_plot_shape_imp2 <- scatter_plot_shape_imp1 +
scale_shape_discrete(name="Years") +
scale_color_discrete(name="Area")

print(scatter_plot_shape_imp2)
```



Improvement of a plot: plot title

```
scatter_plot_shape_imp3 <- scatter_plot_shape_imp2 +
 ggtitle("Biomass of species X and average net depth")

print(scatter_plot_shape_imp3)
```

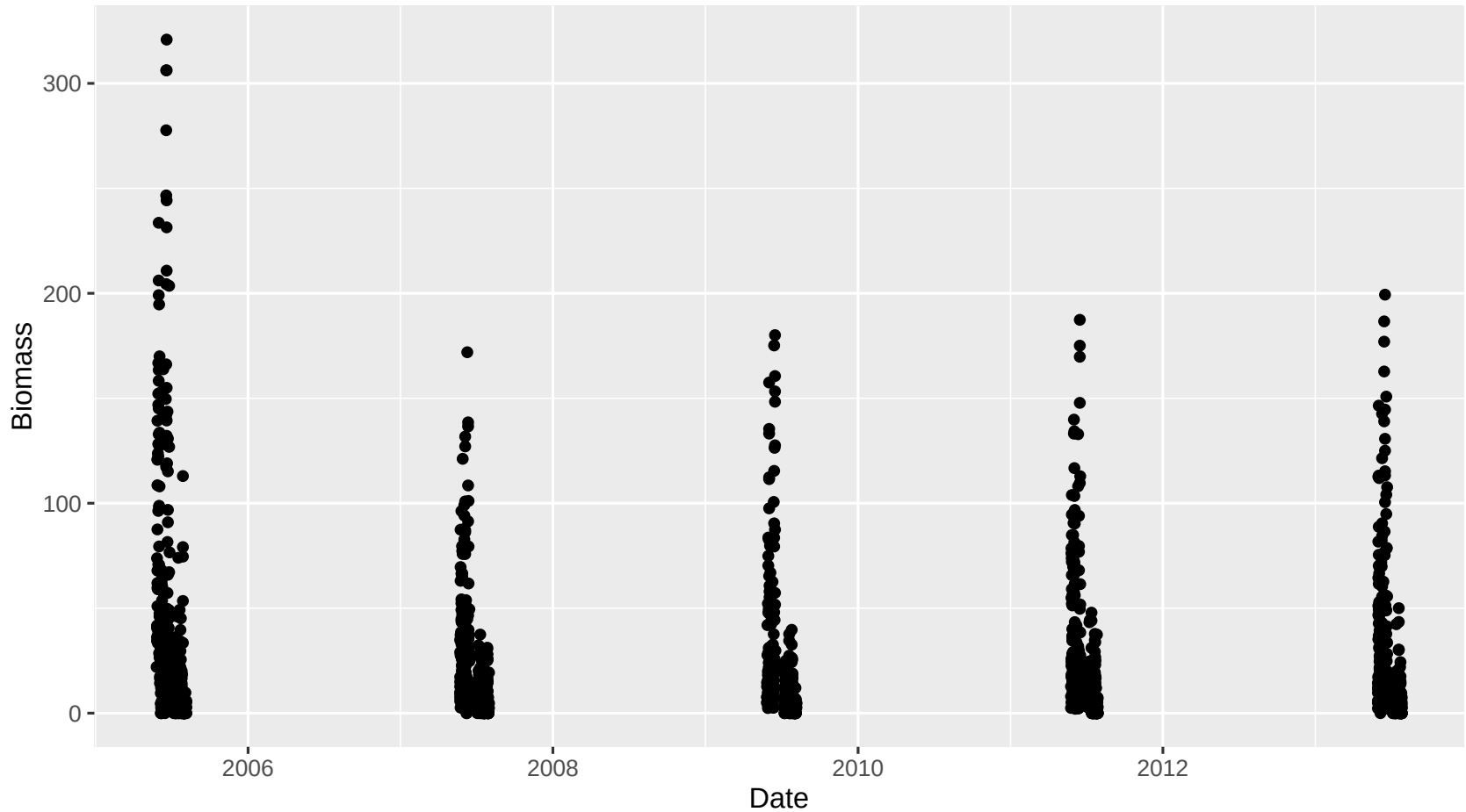


A quick overview of the ggplot2 types

-

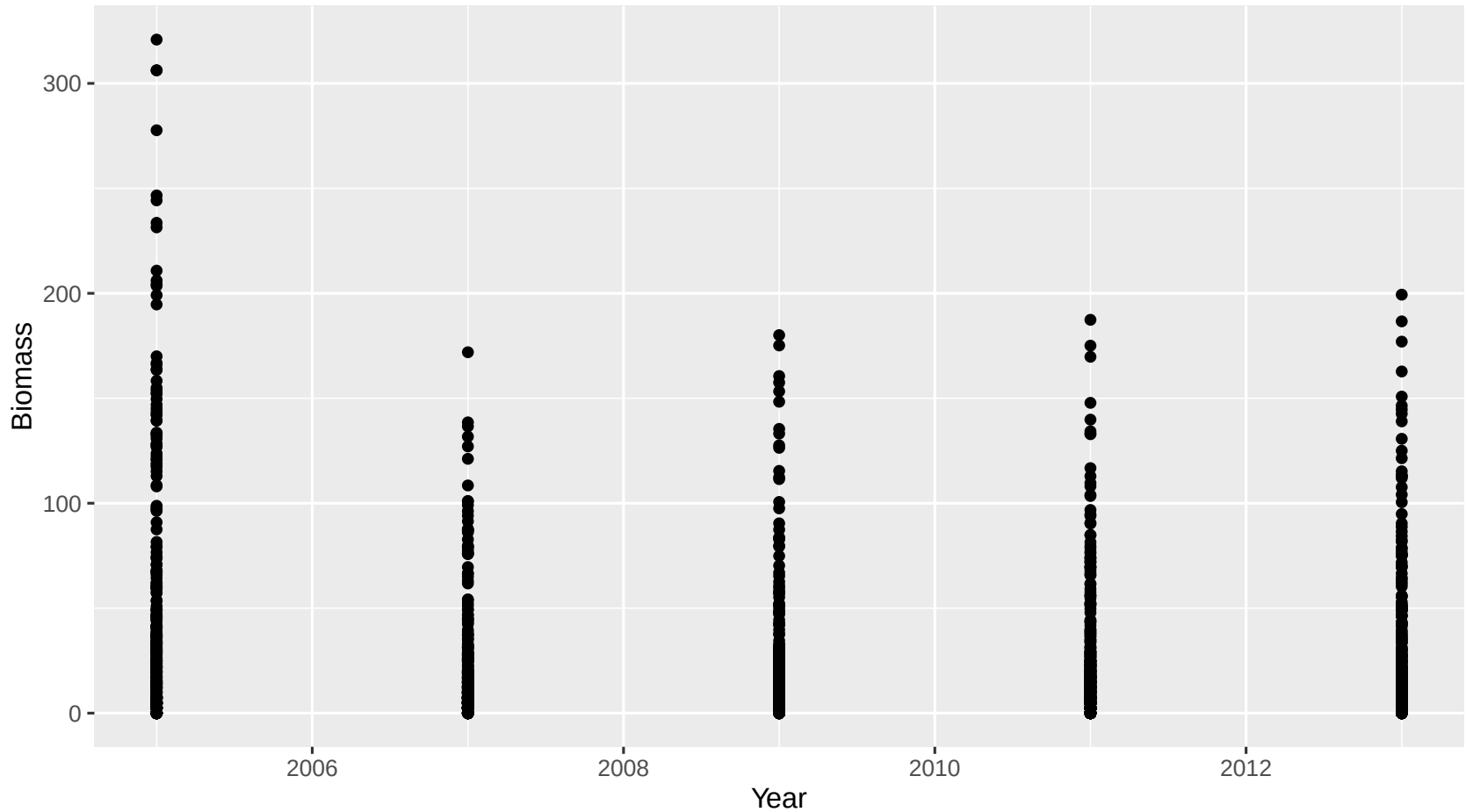
Time series

```
scatter_TSplot <- ggplot(data=df_data, aes(x=Date, y=Biomass)) +
                  geom_point()
print(scatter_TSplot)
```



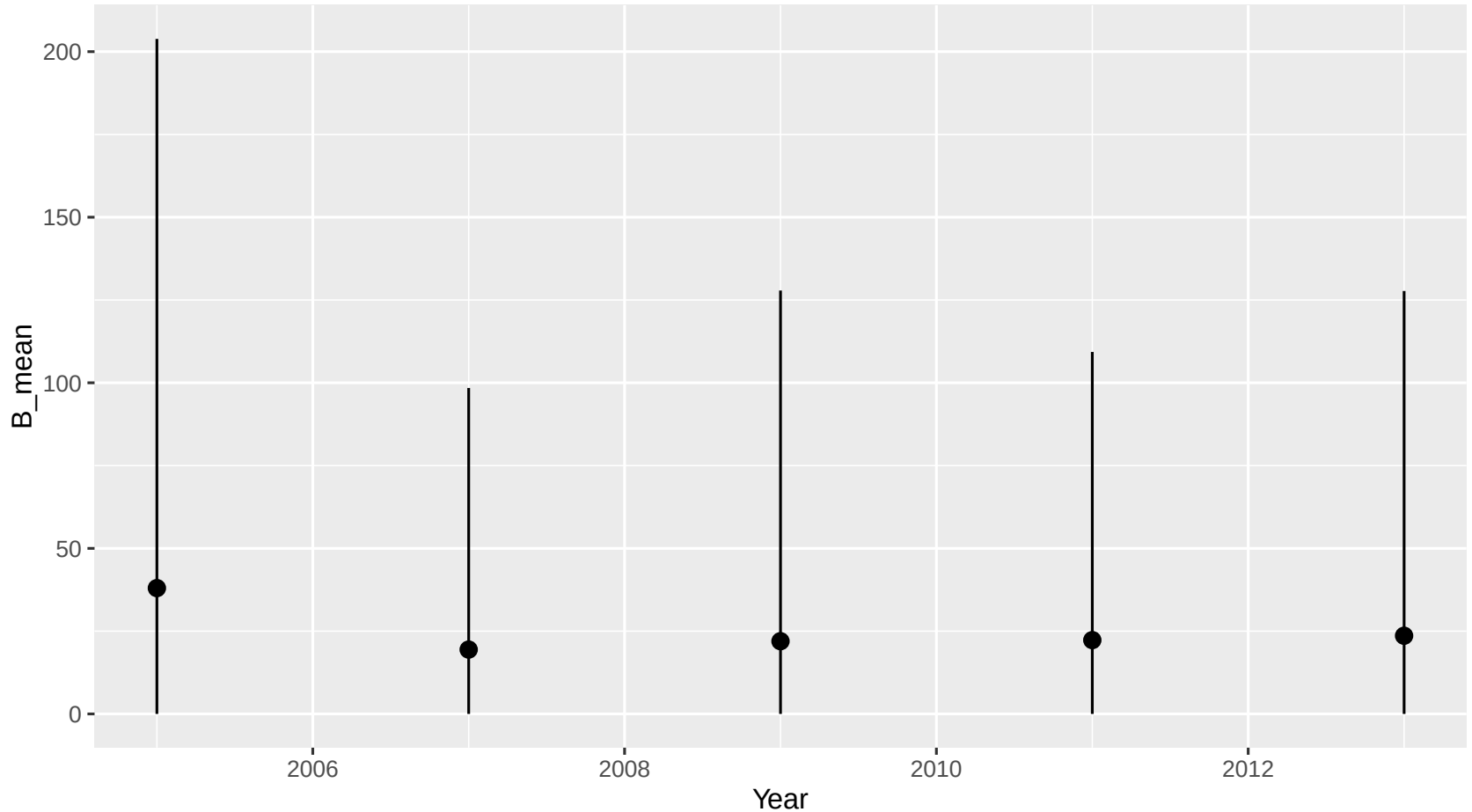
Time series

```
scatter_TSplot_year <- ggplot(data=df_data, aes(x=Year, y=Biomass)) +
                       geom_point()
print(scatter_TSplot_year)
```



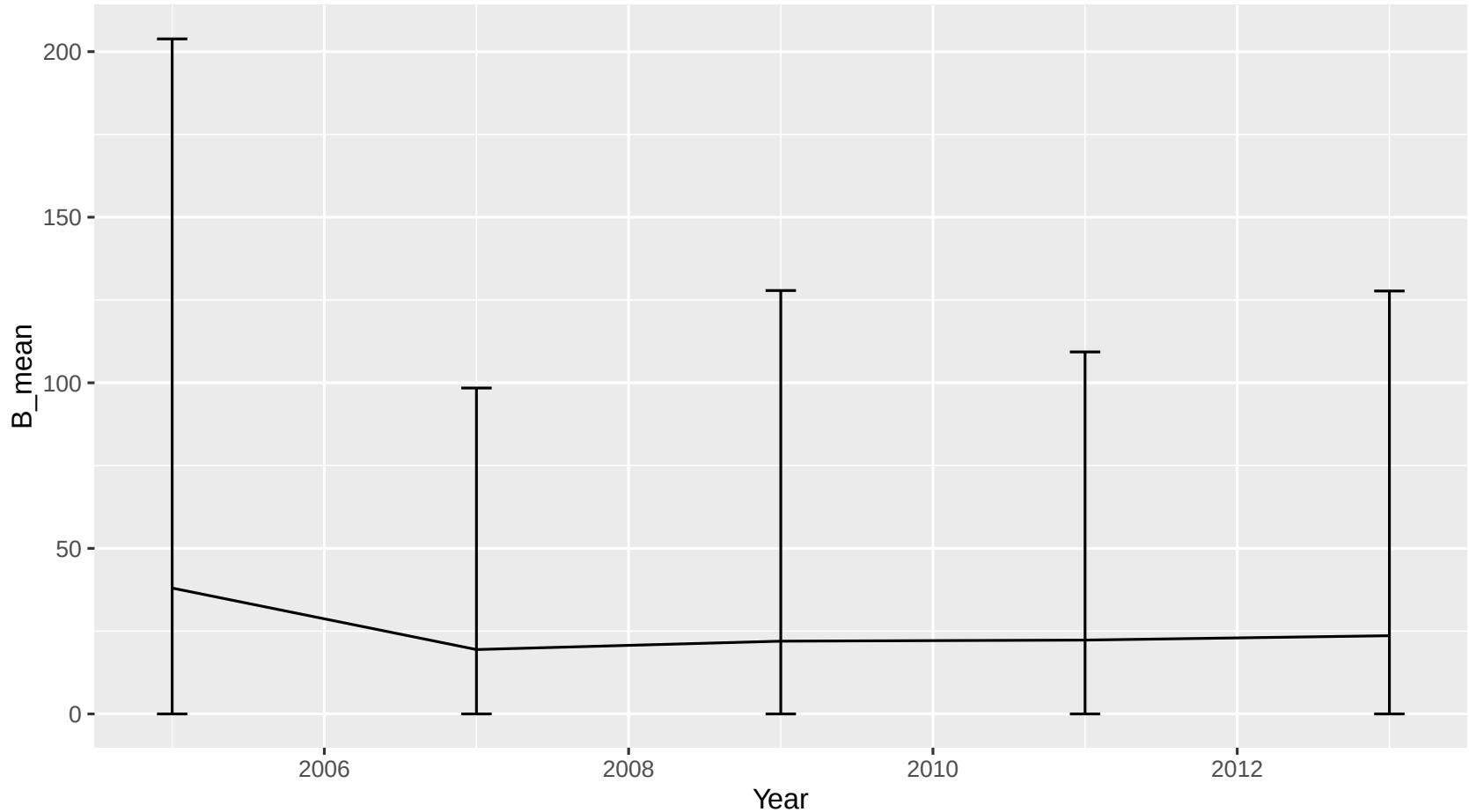
Time series with error bars

```
scatter_TSplot_i95 <- ggplot(data=df_data_summary, aes(x=Year, y=B_mean)
            geom_point() +
            geom_pointrange(aes(ymin = B_q025, ymax = B_q975))
print(scatter_TSplot_i95)
```



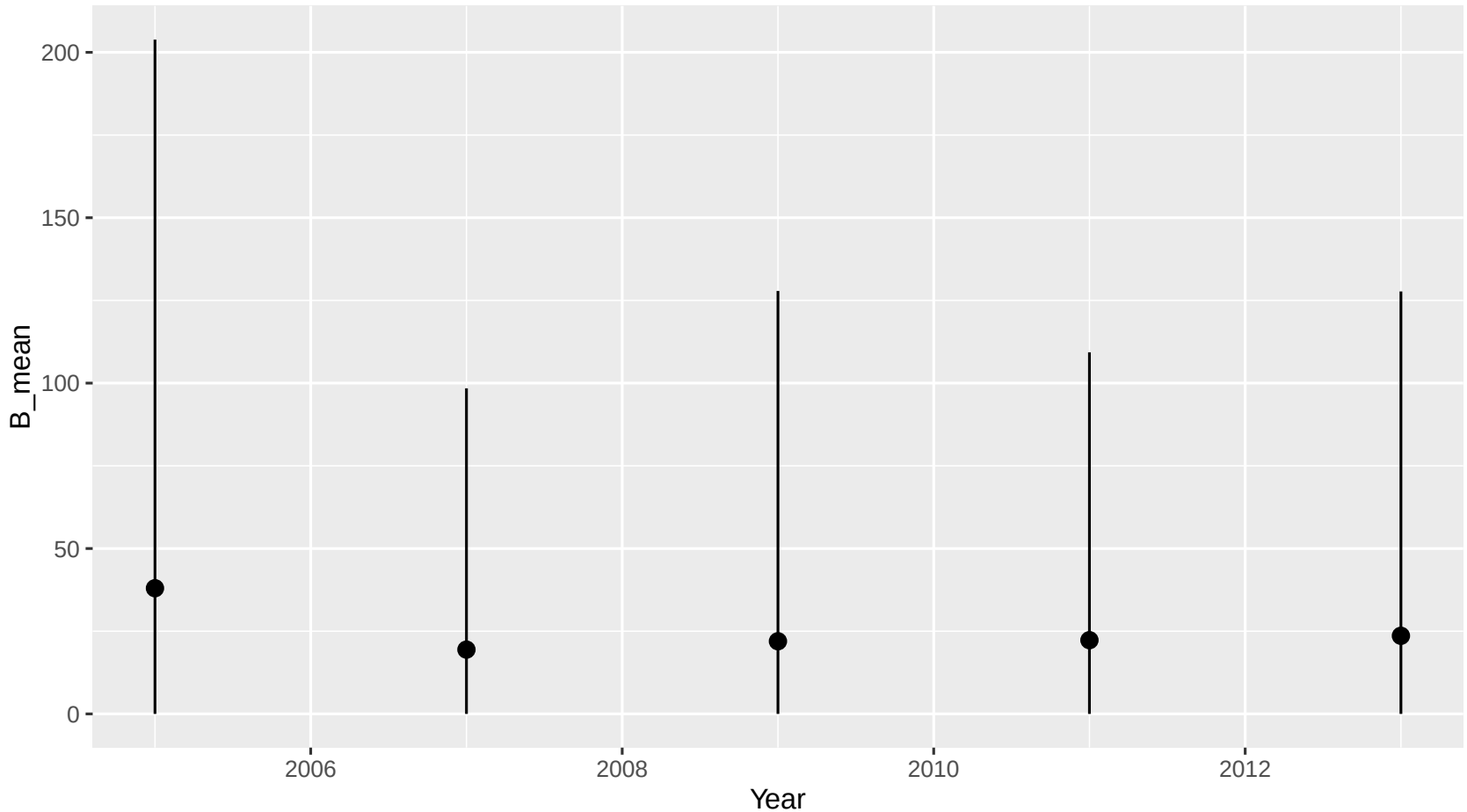
Time series with error bars

```
scatter_TSplot_errori95 <- ggplot(data=df_data_summary, aes(x=Year, y=B_
            geom_line() +
            geom_errorbar(aes(ymin = B_q025, ymax = B_q975), width = 0.2
print(scatter_TSplot_errori95)
```



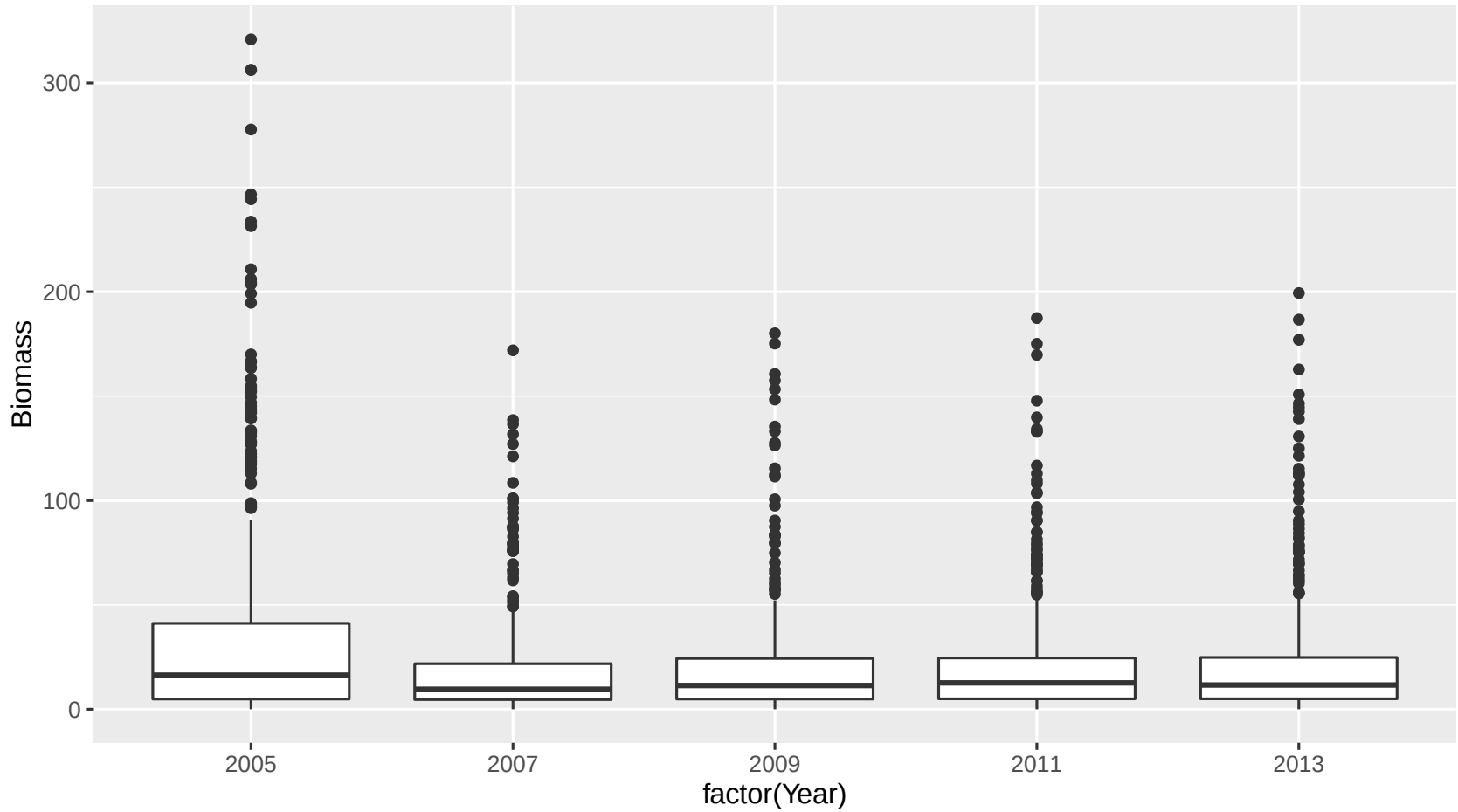
Boxplot

```
   boxplot_TS <- ggplot(data=df_data, aes(x=Year, y=Biomass)) +
geom_boxplot()
   print(scatter_TSplot_i95)
```



Boxplot with Year as a factor

```
boxplot_TSf <- ggplot(data=df_data, aes(x=factor(Year), y=Biomass)) +
  geom_boxplot()
print(boxplot_TSf)
```



Boxplot with Year as a factor

```
boxplot_TS_AREA <- ggplot(data=df_data, aes(x=factor(Year), y=Biomass, col
  geom_boxplot()
print(boxplot_TS_AREA)
```