Taking ggplot2 beyond single plots: Maximizing information transfer

Matthew Castelo, MD
PhD Candidate, IHPME

February, 2022

Matthew Castelo, MD PhD Candidate, IHPME

Overview

Disclosure

- I have no conflicts of interest to disclose
- Please contact me with any questions (matthew.castelo@medportal.ca)

Motivation

I already use ggplot2, why go further?

- Amount of information in a single plot is limited
- Typically allowed 5-8 tables and figures in a manuscript
- Maximize what you can convey in a limited space

Goals

1. Review structure of a ggplot
2. Using aesthetics and annotations
3. Simple and complex faceting
4. Arranging panels
5. Inset tables

Libraries

```
library(tidyverse)
library(ggsci)
library(ggpubr)
library(cowplot)
library(flextable)
library(gtsummary)
library(grid)
```

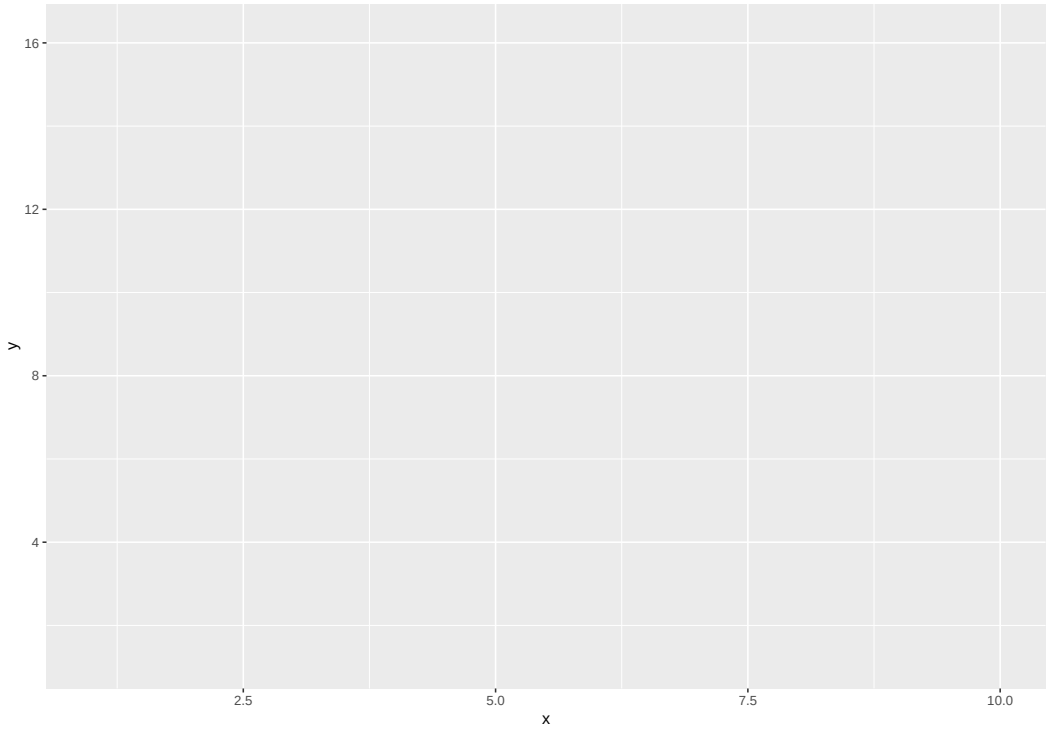
Review of ggplot2

Three steps

1. Map variables in dataset to axes and aesthetics
2. Add layers
3. Customize

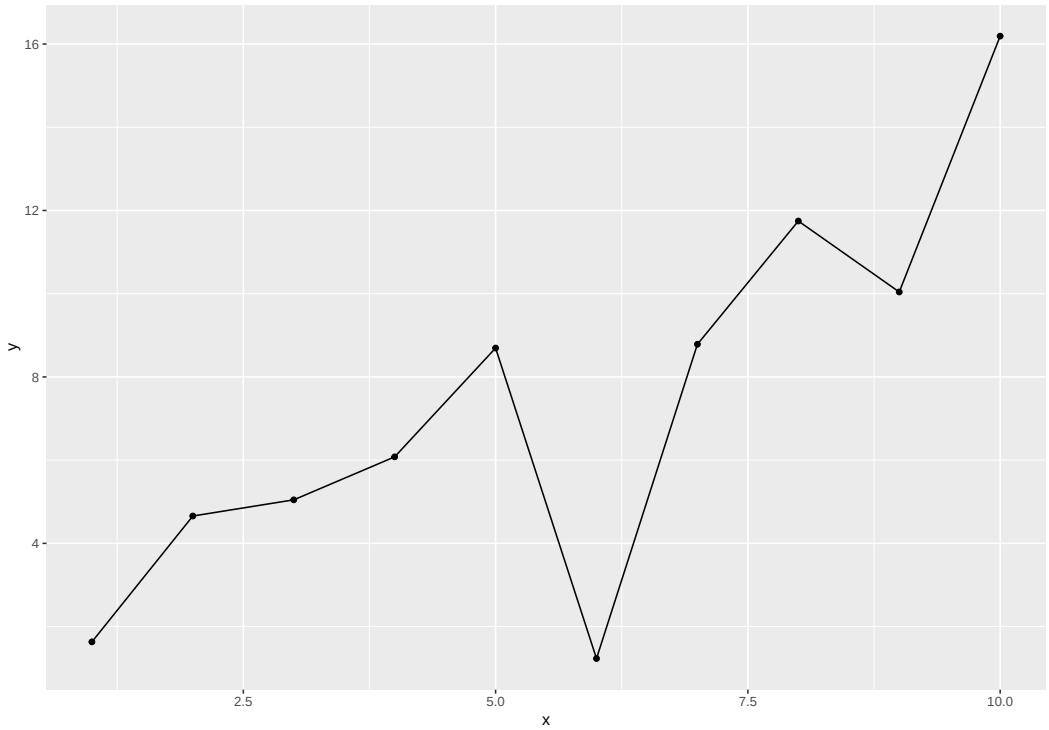
1. Map to axes and aesthetics

```
dataset %>%
  ggplot(aes(x = x, y = y))
```



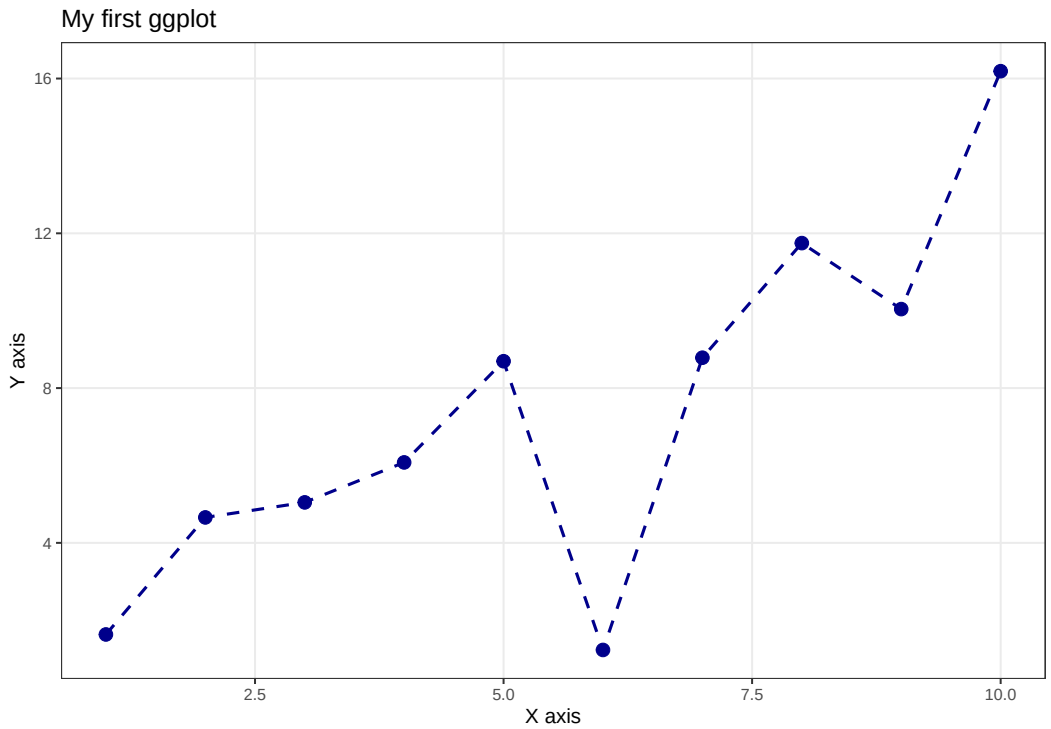
2. Add layers

```
dataset %>%
  ggplot(aes(x = x, y = y))+
  geom_point()+
  geom_line()
```



3. Customize

```
dataset %>%
  ggplot(aes(x = x, y = y))+
  geom_point(color = "darkblue", size = 4)+
  geom_line(color = "darkblue", linetype = 2, size = 1)+
  labs(x = "X axis",
       y = "Y axis",
       title = "My first ggplot")+
  theme_bw(base_size = 14)+
  theme(panel.grid.minor = element_blank())
```



Remember to save!

```
ggsave("my_plot.png", dpi = 300, dev = "png",
       height = 15, width = 15, units = "cm")
```

Also possible to save as PDF, TIFF, SVG, etc

Taking advantage of aesthetics and annotations

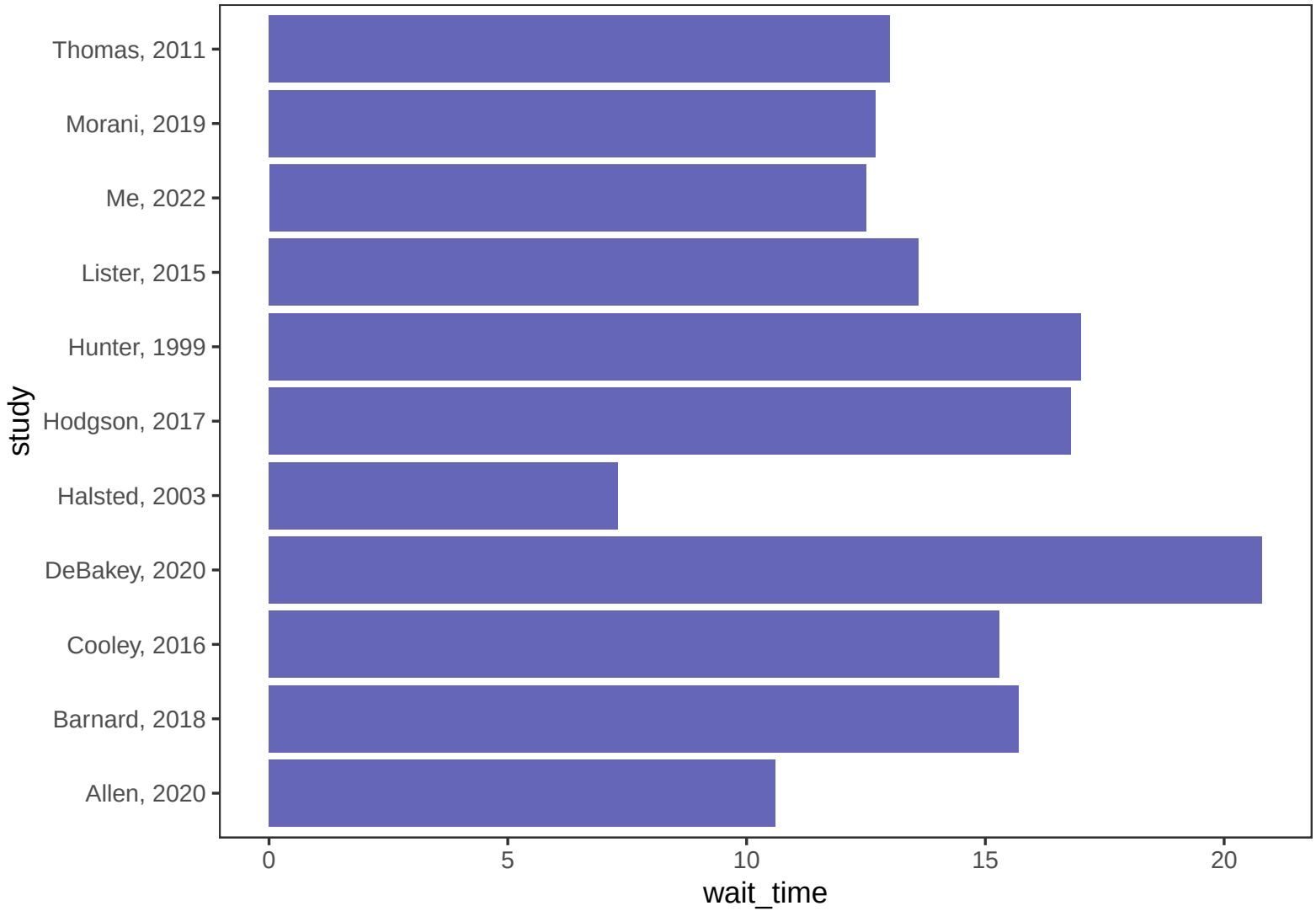
Before splitting a plot into facets or panels, first use aesthetics to show more information in a single plot

Motivating example

You measure wait time in cancer patients, and want to compare your results to the literature

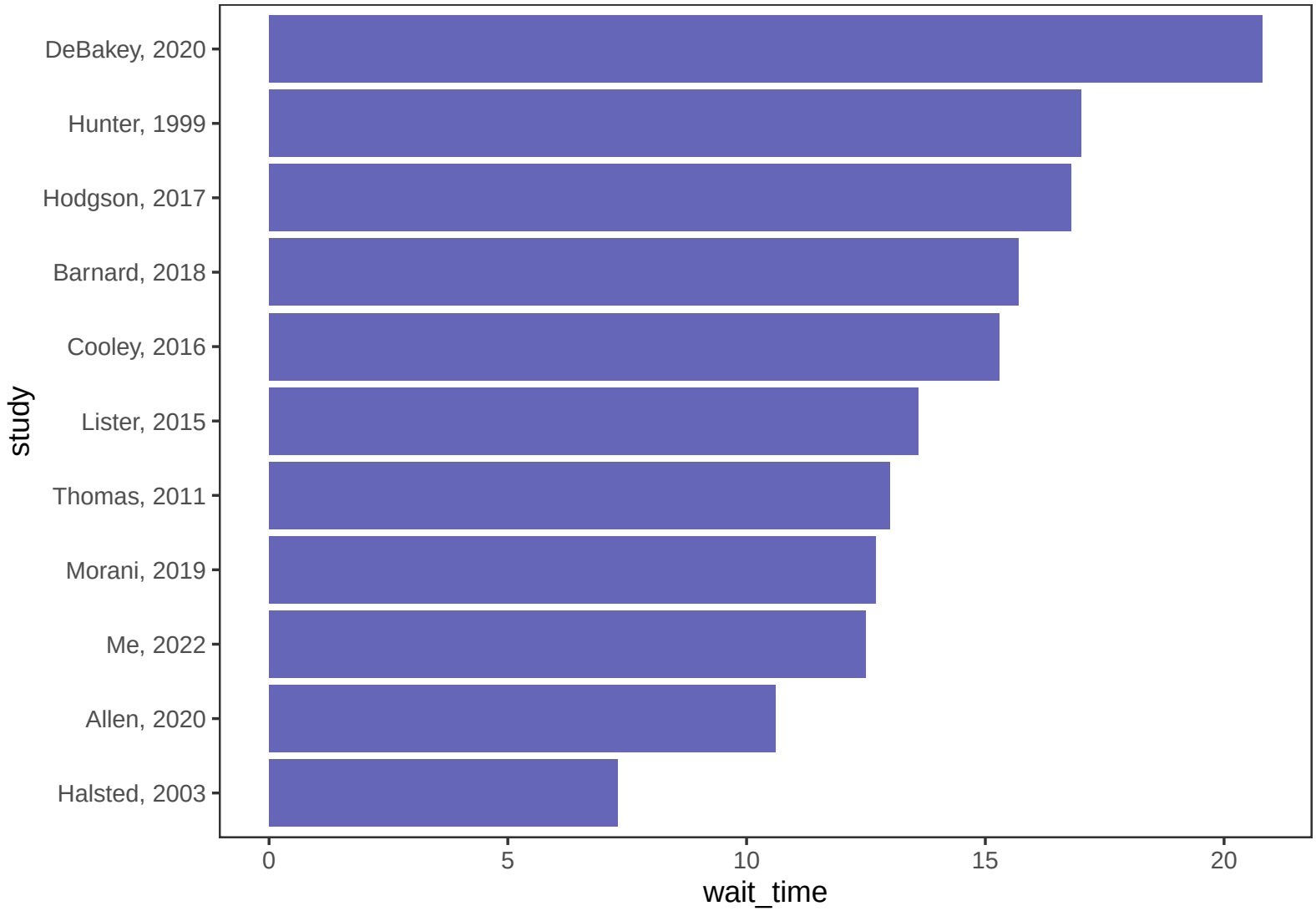
study	wait_time	time_males	time_females	measure	n
Me, 2022	12.5	8.0	16.5	Median	58
Cooley, 2016	15.3	14.2	19.5	Median	77
Barnard, 2018	15.7	11.6	22.6	Mean	83
DeBakey, 2020	20.8	16.0	24.7	Mean	82
Hunter, 1999	17.0	14.6	20.9	Median	81
Halsted, 2003	7.3	5.2	11.1	Mean	106

```
wait %>%
  ggplot(aes(y = study, x = wait_time))+
  geom_col(fill = "darkblue", alpha = 0.6)+
  my_theme
```



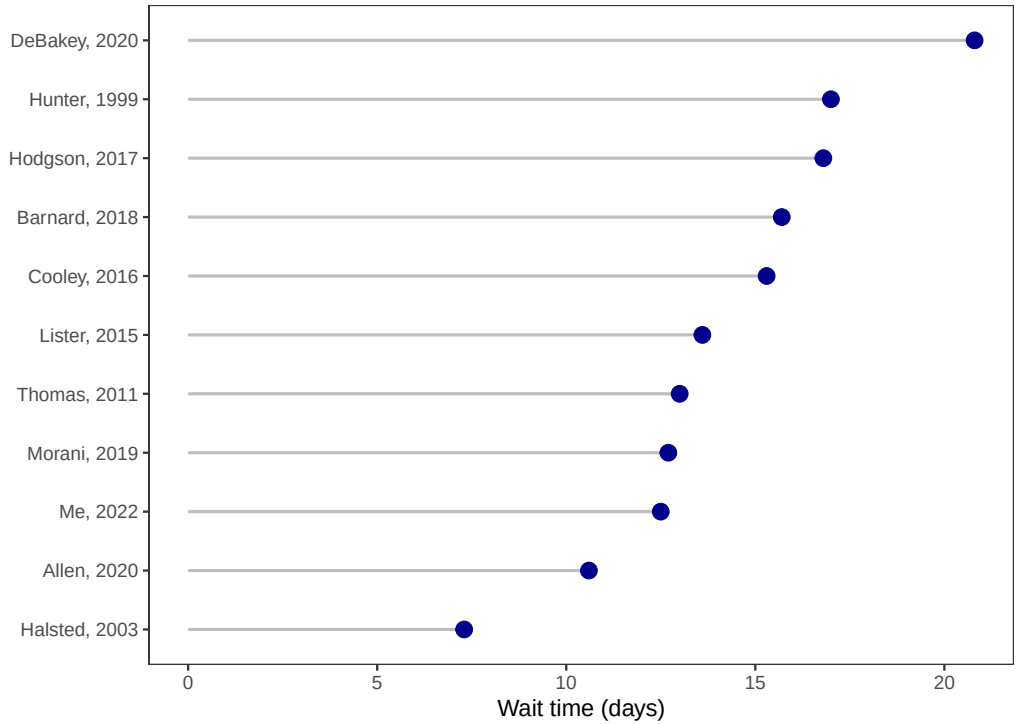
Reorder study by another variable

```
wait <- wait %>%
  mutate(study = fct_reorder(study, wait_time))
```



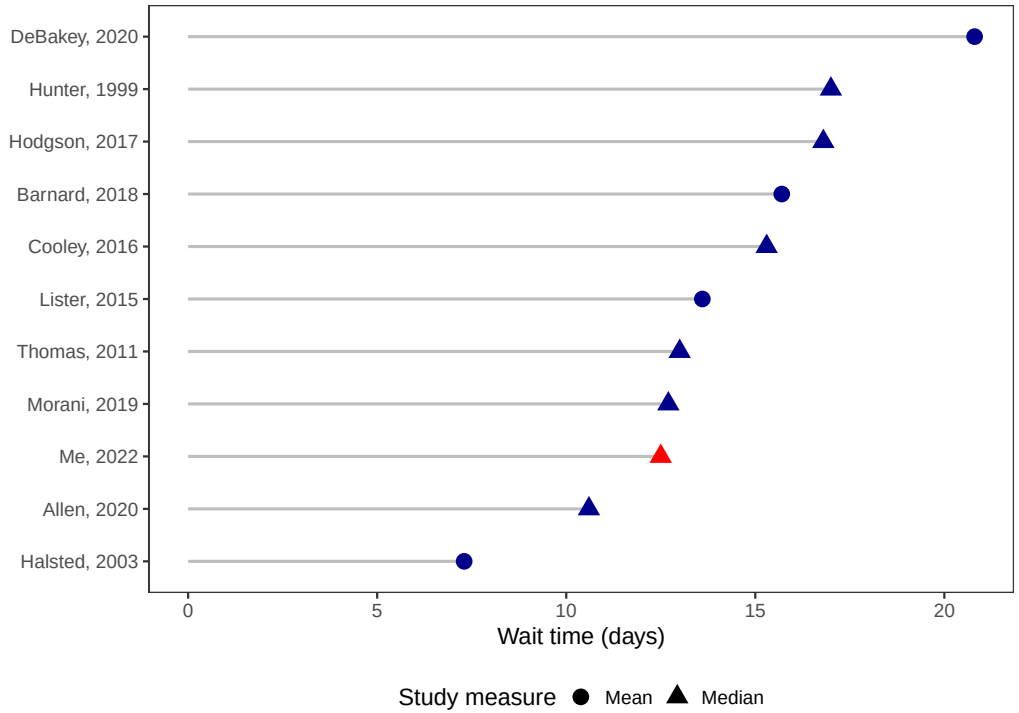
Lollipop chart and customize

```
wait %>%
  ggplot(aes(y = study, x = wait_time))+
  geom_segment(aes(yend = study,
                   x = 0, xend = wait_time),
               size = 1, color = "grey")+
  geom_point(color = "darkblue", size = 5)+
  labs(x = "Wait time (days)", y = "")+
  my_theme
```



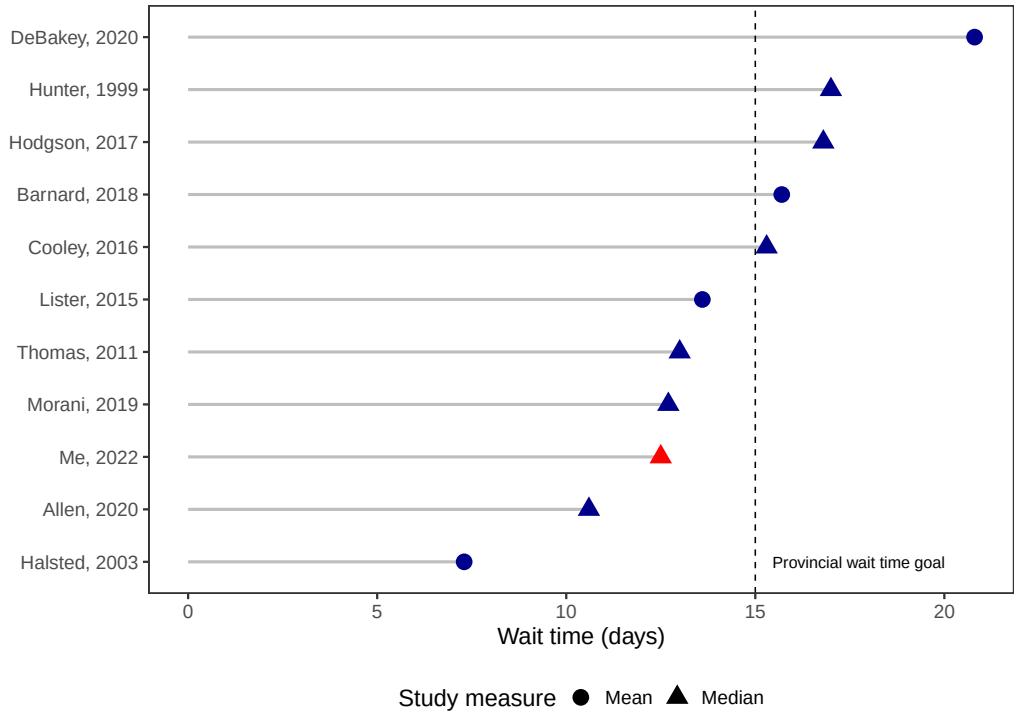
Show measure by shape and highlight our study

```
wait %>%
  ggplot(aes(y = study, x = wait_time, shape = measure,
             color = ifelse(study == "Me, 2022", "1", "2")))+
  scale_color_manual(values = c("red", "darkblue"),
                     guide = "none")+
```



Annotations

```
p+
  geom_vline(xintercept = 15, linetype = 2)+
  annotate(geom = "text",
           x = 15, y = "Halsted, 2003", hjust = -0.1,
           label = "Provincial wait time goal")
```

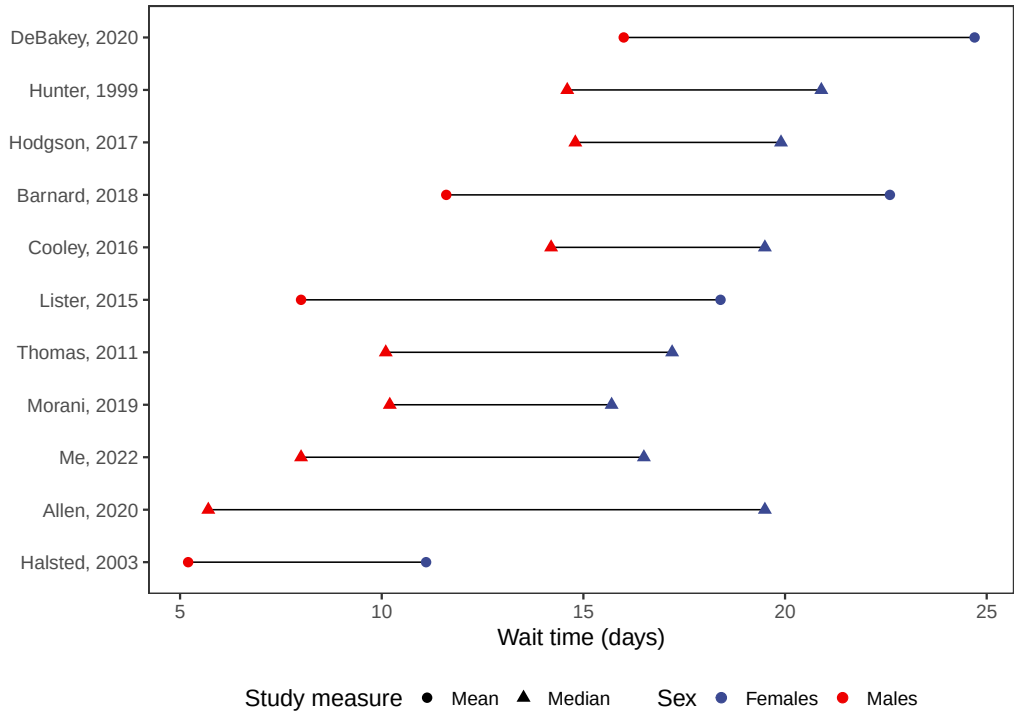


Can show data for males and females

```
wait_long <- wait %>%
  pivot_longer(cols = c(time_males,
                        time_females),
               names_to = "Sex") %>%
  mutate(Sex = str_to_title(str_replace_all(Sex, "time_", "")))
```

study	measure	Sex	value
Me, 2022	Median	Males	8.0
Me, 2022	Median	Females	16.5
Cooley, 2016	Median	Males	14.2
Cooley, 2016	Median	Females	19.5
Barnard, 2018	Mean	Males	11.6
Barnard, 2018	Mean	Females	22.6

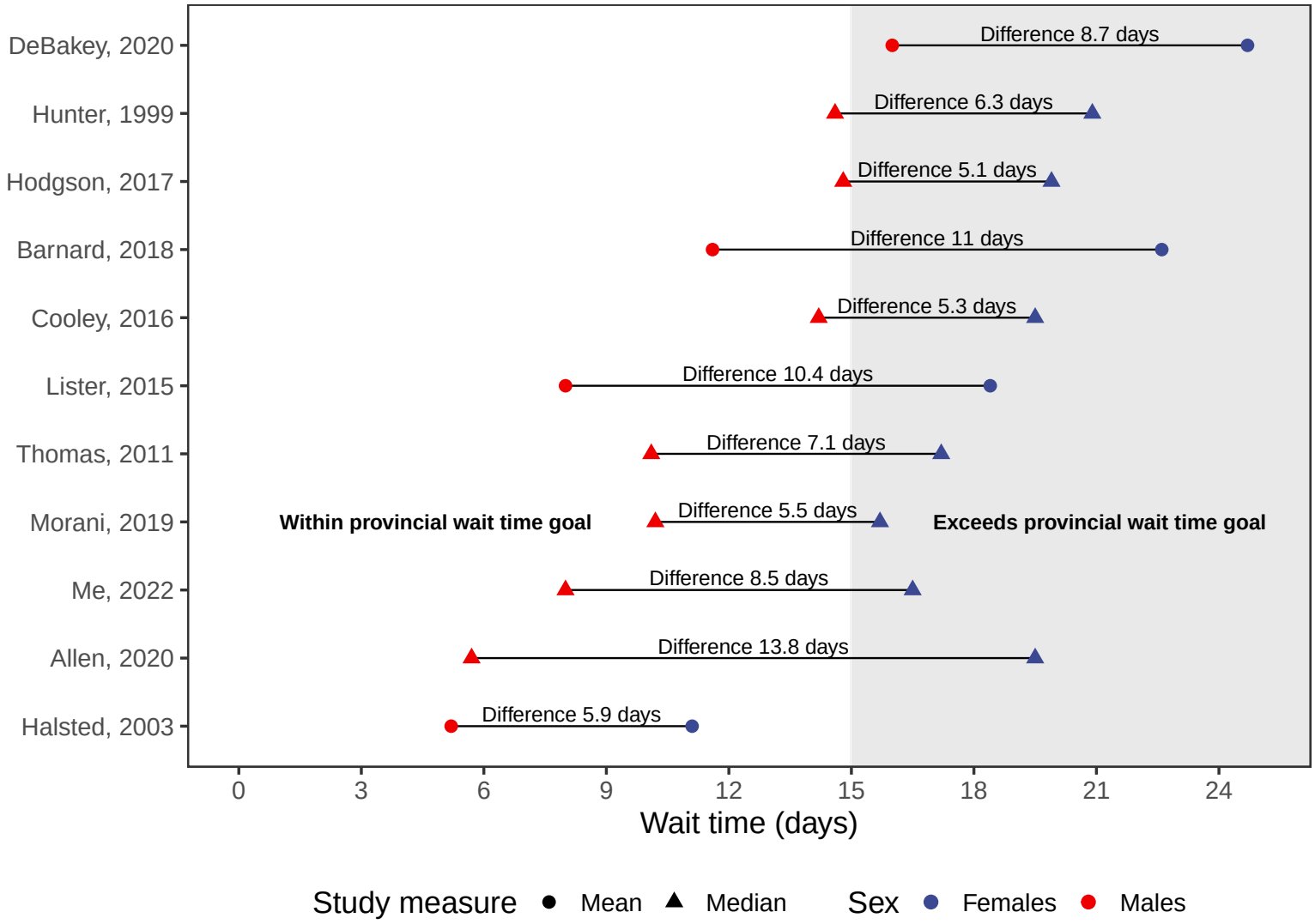
```
wait %>%
  ggplot(aes(y = study, yend = study,
             x = time_males, xend = time_females))+
  geom_segment()+
  geom_point(aes(y = study, x = value, shape = measure,
                 color = Sex),
             size = 3, data = wait_long, inherit.aes = FALSE)+
  scale_color_aaas()+
  labs(y = "", x = "Wait time (days)",
       shape = "Study measure")+
  my_theme
```



Add text as a layer

```
wait %>%
  mutate(text_label = paste0("Difference ",
                             time_females - time_males,
                             " days")) %>%
  ggplot(aes(y = study, yend = study,
             x = time_males, xend = time_females))+
  geom_rect(aes(xmin = 15, xmax = Inf,
                ymax = Inf, ymin = -Inf),
            color = "grey95", alpha = 0.01)+
  geom_segment()+
  geom_point(aes(y = study, x = value, shape = measure,
                 color = Sex),
             size = 3, data = wait_long, inherit.aes = FALSE)+
  geom_text(aes(x = (time_males + time_females) / 2,
                label = text_label),
            vjust = -0.3, hjust = 0.5)+
  scale_x_continuous(limits = c(0, 25),
                     breaks = seq(0, 25, 3))+
  annotate(geom = "text", x = 17, y = "Morani, 2019",
           hjust = 0,
           label = "Exceeds provincial wait time goal",
           fontface = "bold")+
  annotate(geom = "text", x = 1, y = "Morani, 2019",
           hjust = 0,
           label = "Within provincial wait time goal",
           fontface = "bold")+
  scale_color_aaas()+
  labs(y = "", x = "Wait time (days)",
       shape = "Study measure")+
  my_theme
```

Matthew Castelo, MD PhD Candidate, IHPME



Using facets

Facets create sub-plots stratified by a grouping variable

Motivating example

You administer a survey before and after the pandemic. Nine questions are grouped into four categories.

The data are organized as two spreadsheets (before and after the pandemic)

name	value	time
question_1	3.8	Pre-pandemic
question_2	0.0	Pre-pandemic
question_3	1.6	Pre-pandemic
question_4	5.8	Pre-pandemic
question_5	0.3	Pre-pandemic
question_6	4.5	Pre-pandemic

name	value	time
question_1	0.2	Post-pandemic
question_2	4.5	Post-pandemic
question_3	1.6	Post-pandemic
question_4	3.2	Post-pandemic
question_5	5.7	Post-pandemic
question_6	3.6	Post-pandemic

We have a key with the question names and categories

quest_num	category	question
question_1	Dejection	I have been worried
question_2	Dejection	I have been uneasy
question_3	Dejection	I have felt sad
question_4	Anxiety	I have felt scared
question_5	Anxiety	I have been nervous
question_6	Anxiety	I have felt restless
question_7	Sleep	I have slept badly
question_8	Sleep	I have woken up too early
question_9	Other	I have taken many sick days

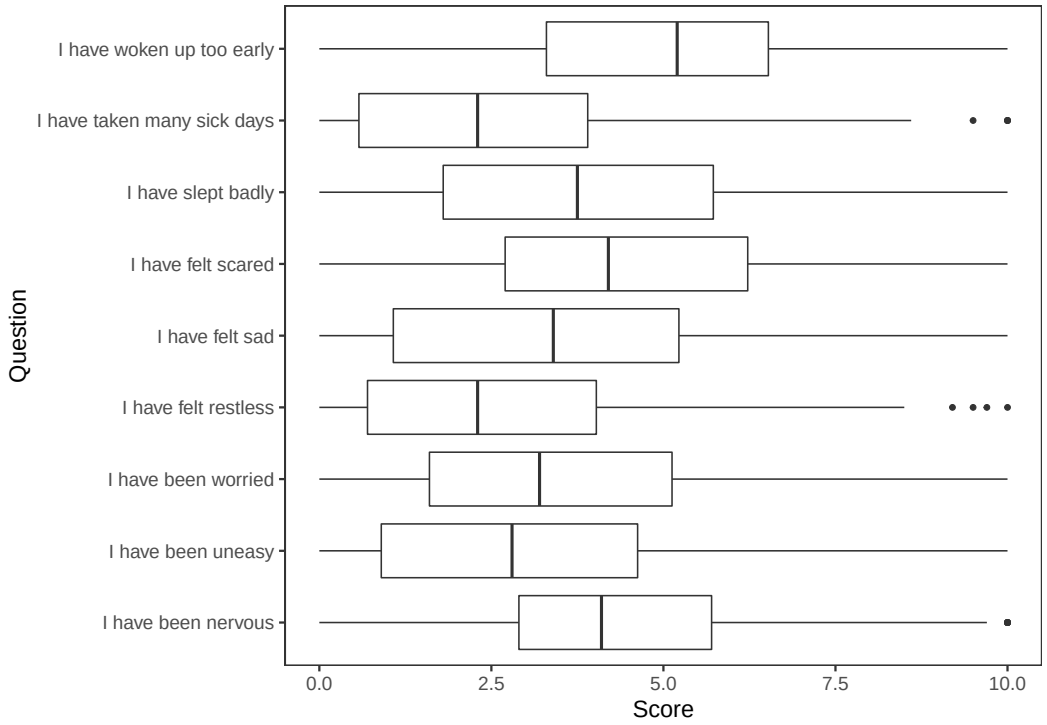
Bind the data in long form for plotting

```
survey <- rbind(before,
                after) %>%
  left_join(key, by = c("name" = "quest_num"))
```

name	value	time	category	question
question_1	3.8	Pre-pandemic	Dejection	I have been worried
question_2	0.0	Pre-pandemic	Dejection	I have been uneasy
question_3	1.6	Pre-pandemic	Dejection	I have felt sad
question_4	5.8	Pre-pandemic	Anxiety	I have felt scared
question_5	0.3	Pre-pandemic	Anxiety	I have been nervous
question_6	4.5	Pre-pandemic	Anxiety	I have felt restless

Simple boxplot

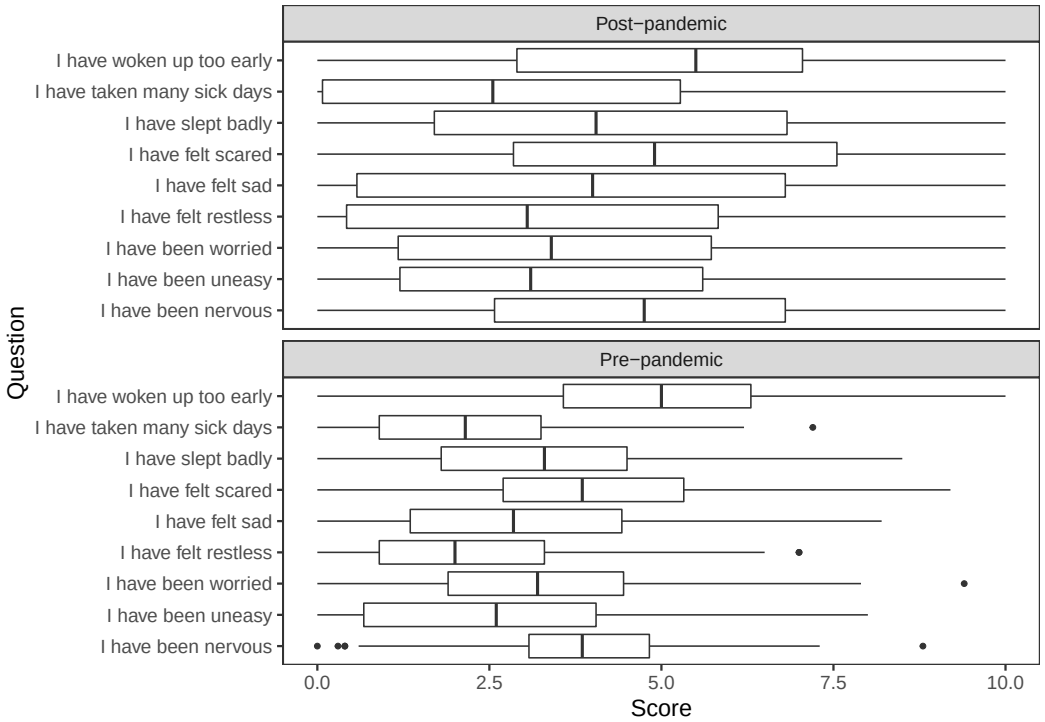
```
survey %>%
  ggplot(aes(y = question, x = value))+
  geom_boxplot()+
  labs(x = "Score", y = "Question")+
  my_theme
```



Simple facet

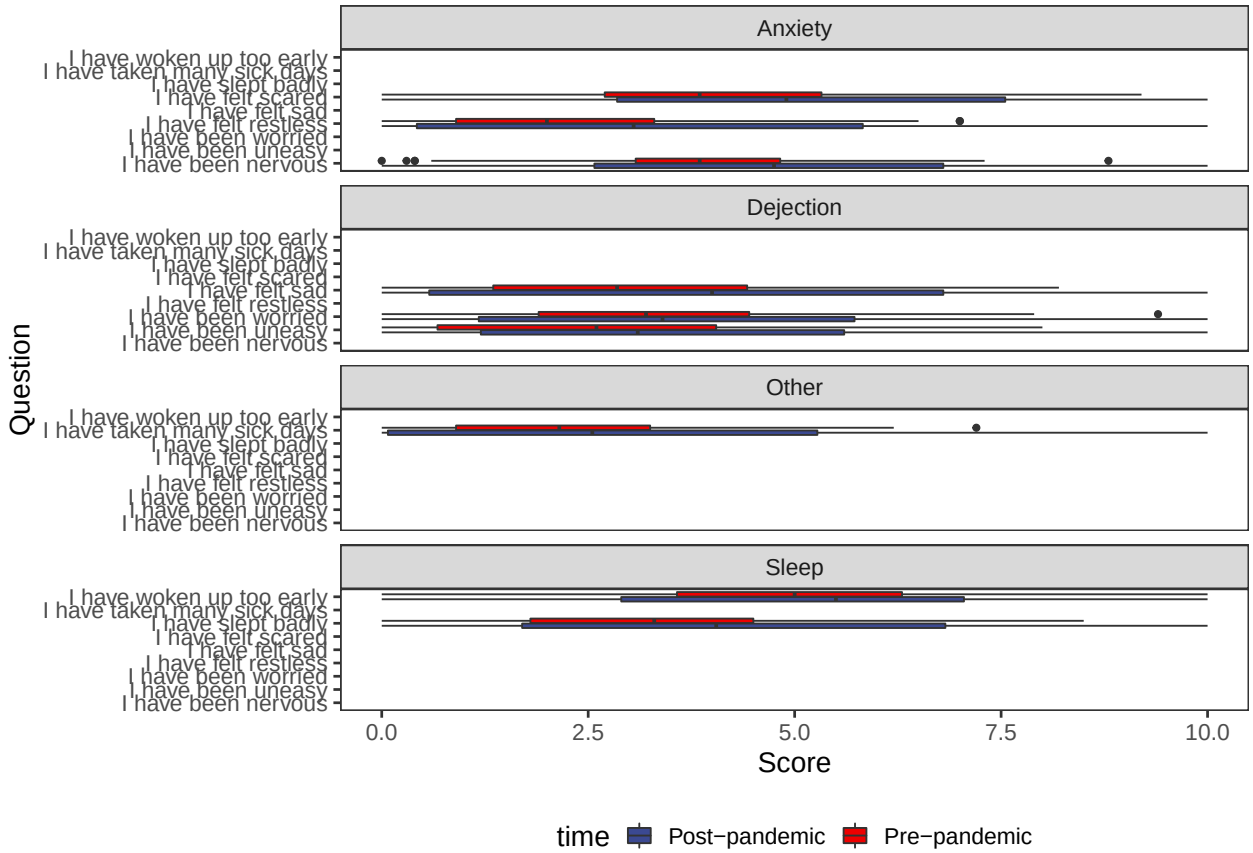
We want to show both time period and category

```
p+
  facet_wrap(~time, ncol = 1)
```



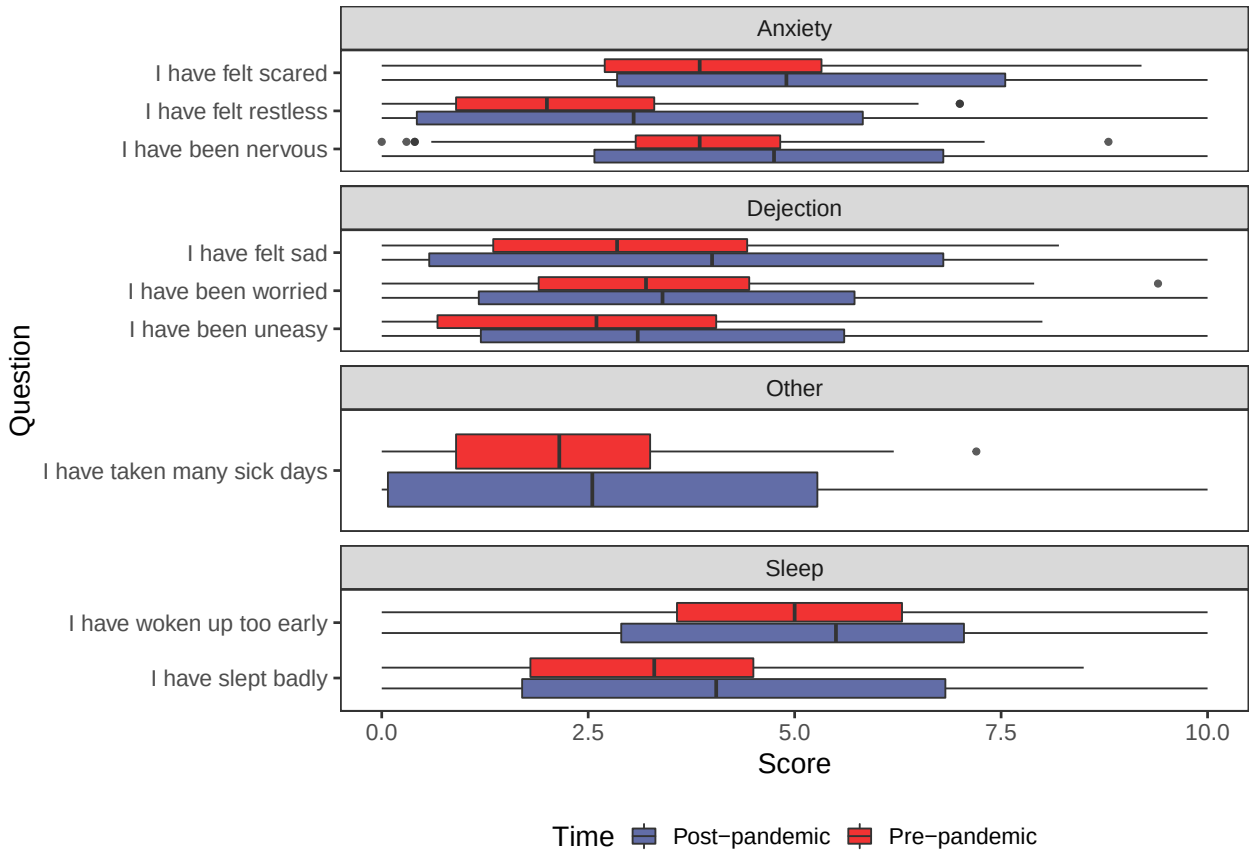
Better to fill by time and facet by category

```
survey %>%
  ggplot(aes(y = question, x = value, fill = time))+
  geom_boxplot()+
  facet_wrap(~category, ncol = 1)+
  labs(x = "Score", y = "Question")+
  scale_fill_aaas()+
  my_theme
```



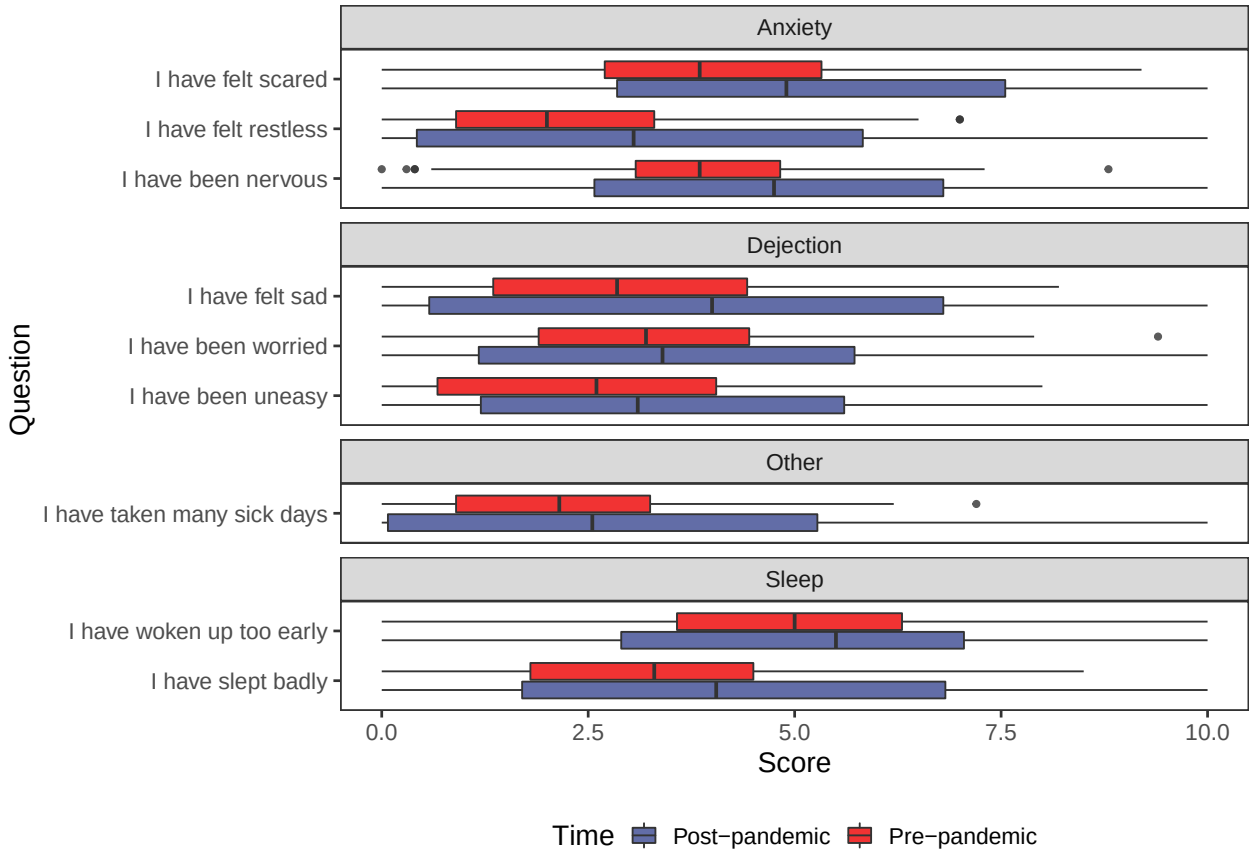
Can set either x or y axis scales to free to remove empty levels

```
p+
  facet_wrap(~category, ncol = 1,
             scales = "free_y")
```



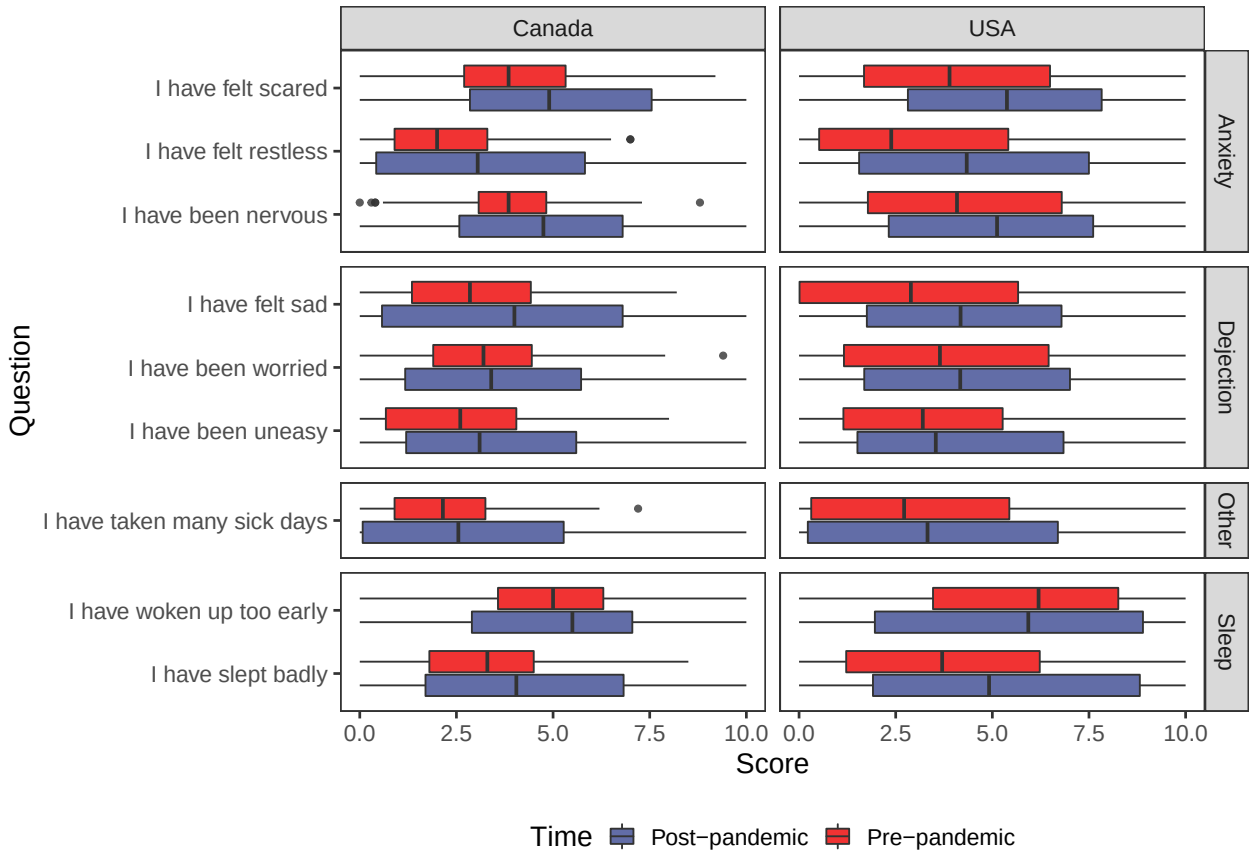
By default, all facets are the same size. The ggforce package can help by automatically shrinking smaller categories

```
p+
  ggforce::facet_col(vars(category),
                     scales = "free_y",
                     space = "fixed",
                     shrink = TRUE)
```



facet_grid() facets by two variables

```
p+
  facet_grid(rows = vars(category),
             cols = vars(country),
             scales = "free",
             space = "free")
```



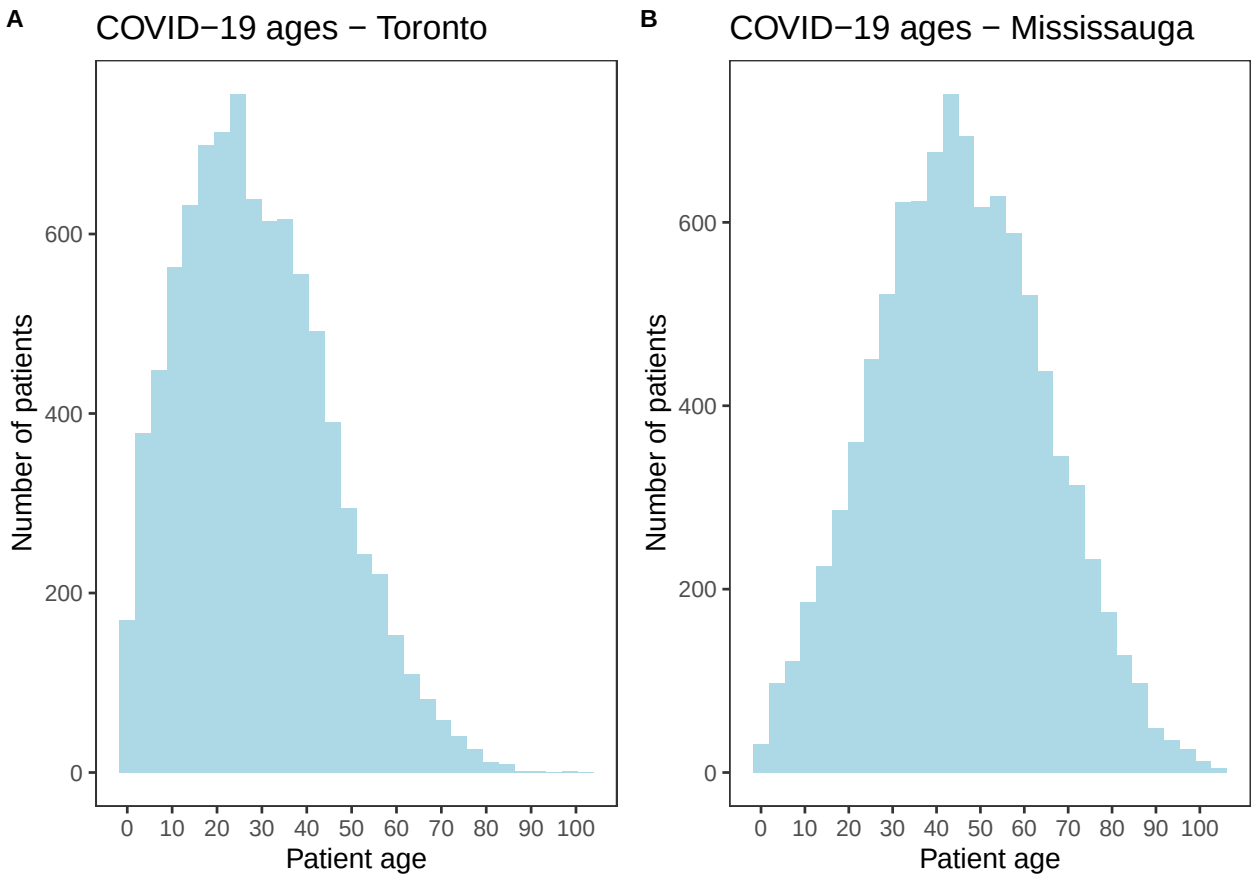
Panelled figures

While facets are useful to organize similar types of data into categories, you are limited in controlling size. Panelled figures can present different chart types and even tables together.

ggarrange()

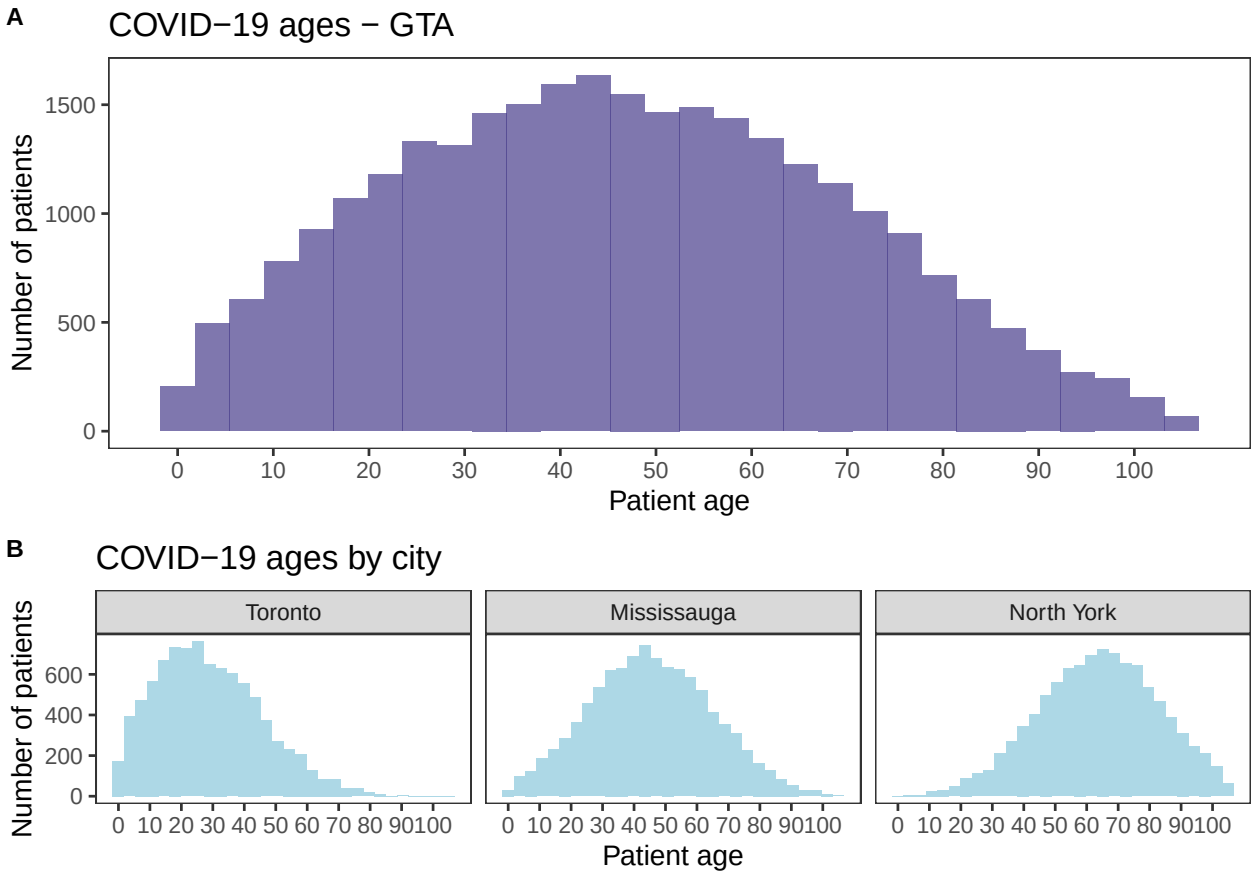
Simply list ggplots inside the function and specify the number of columns/rows

```
ggarrange(plot_1, plot_2,
          nrow = 1,
          labels = "AUTO")
```



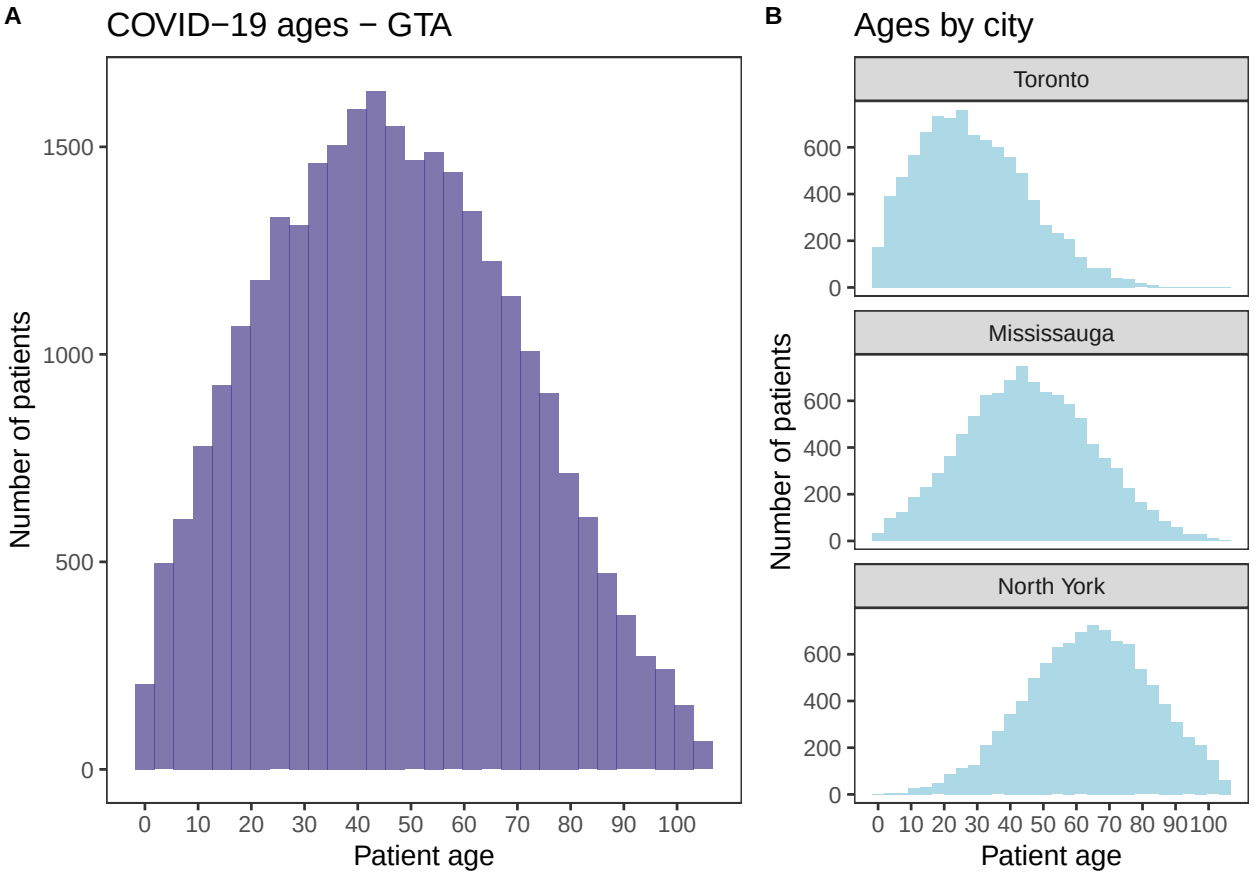
Common to combine faceting and arranging

```
ggarrange(large_plot, faceted_plot,
          ncol = 1, labels = "AUTO",
          heights = c(1.5, 1))
```



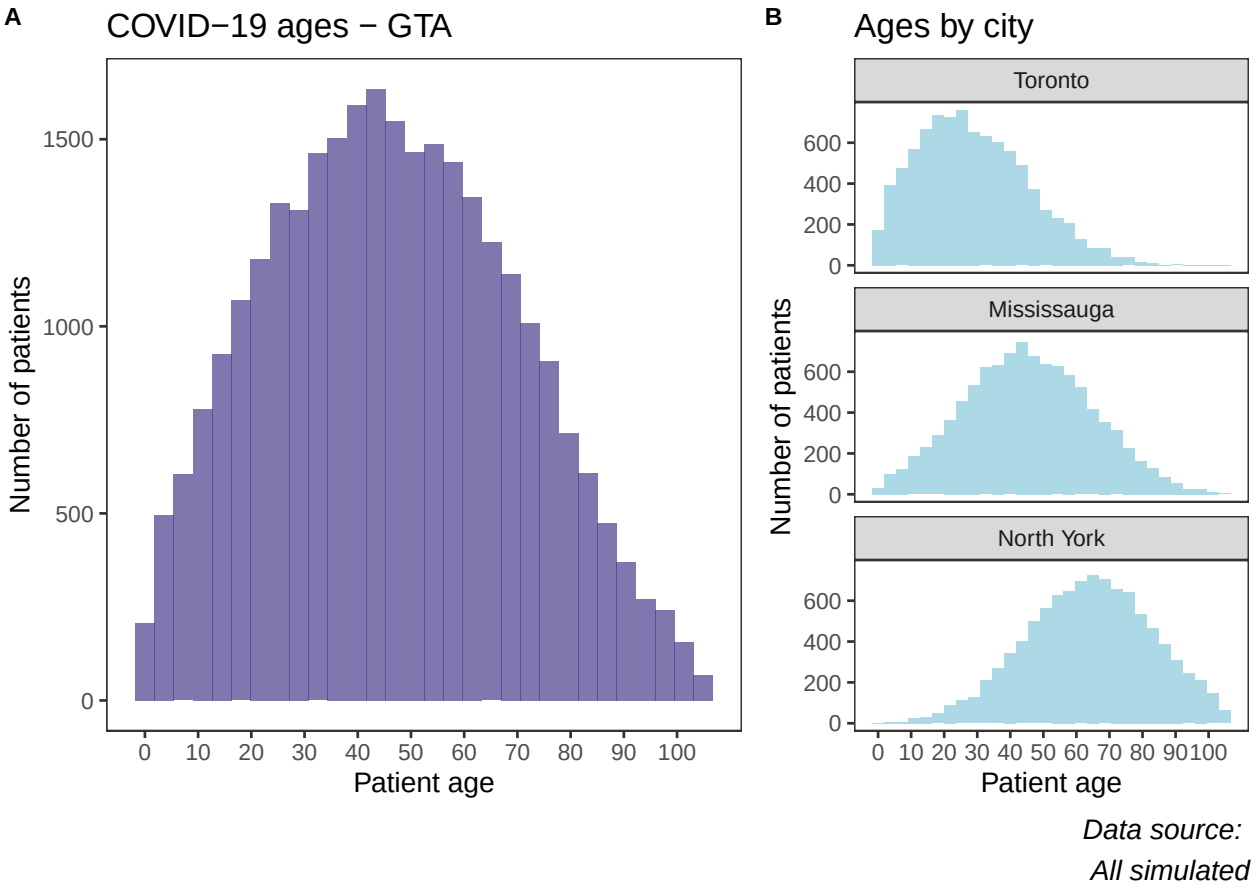
An alternative arrangement

```
ggarrange(large_plot, faceted_plot,
          ncol = 2, labels = "AUTO",
          widths = c(1.5, 1))
```



Add an annotation

```
ggarrange(large_plot, faceted_plot,
          ncol = 2, labels = "AUTO",
          widths = c(1.5, 1)) %>%
  annotate_figure(bottom = text_grob("Data source: \n All simulated",
                                     hjust = 1, x = 0.99,
                                     face = "italic", size = 16))
```



Inset tables

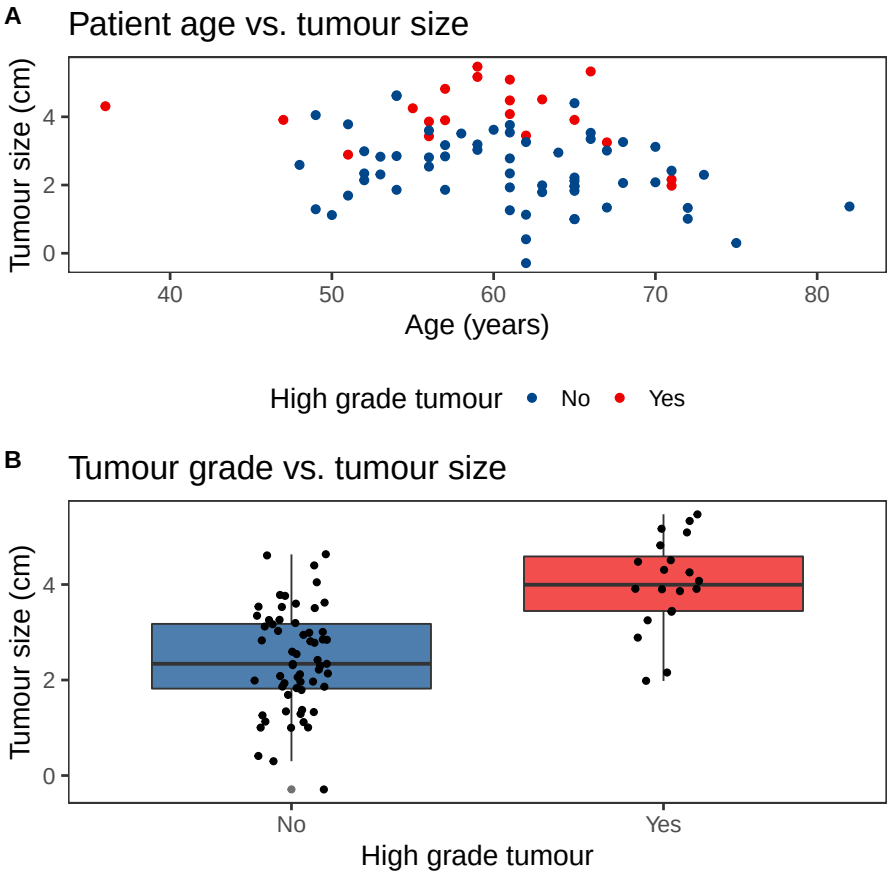
There are several options for inset tables. The flextable package is highly customizable and allows tables to be converted to rasters and ggplots.

Motivating examples

You are studying the effect of tumour grade in cancer patients and wish to present descriptive statistics as a panelled figure.

First arranging a scatterplot and boxplot

```
ggarrange(scatter_plot, boxplot,
          ncol = 1, labels = "AUTO")
```



Creating a summary table with gtsummary and flextable

```
cancer %>%
  tbl_summary(include = -pt_id,
              by = high_grade) %>%
  add_p() %>%
  add_overall() %>%
  bold_p() %>%
  bold_labels() %>%
  as_flex_table() %>%
  add_header_lines("Patient characteristics by high tumour grade") %>%
  theme_vanilla() %>%
  vline(j = 2, part = "all")
```

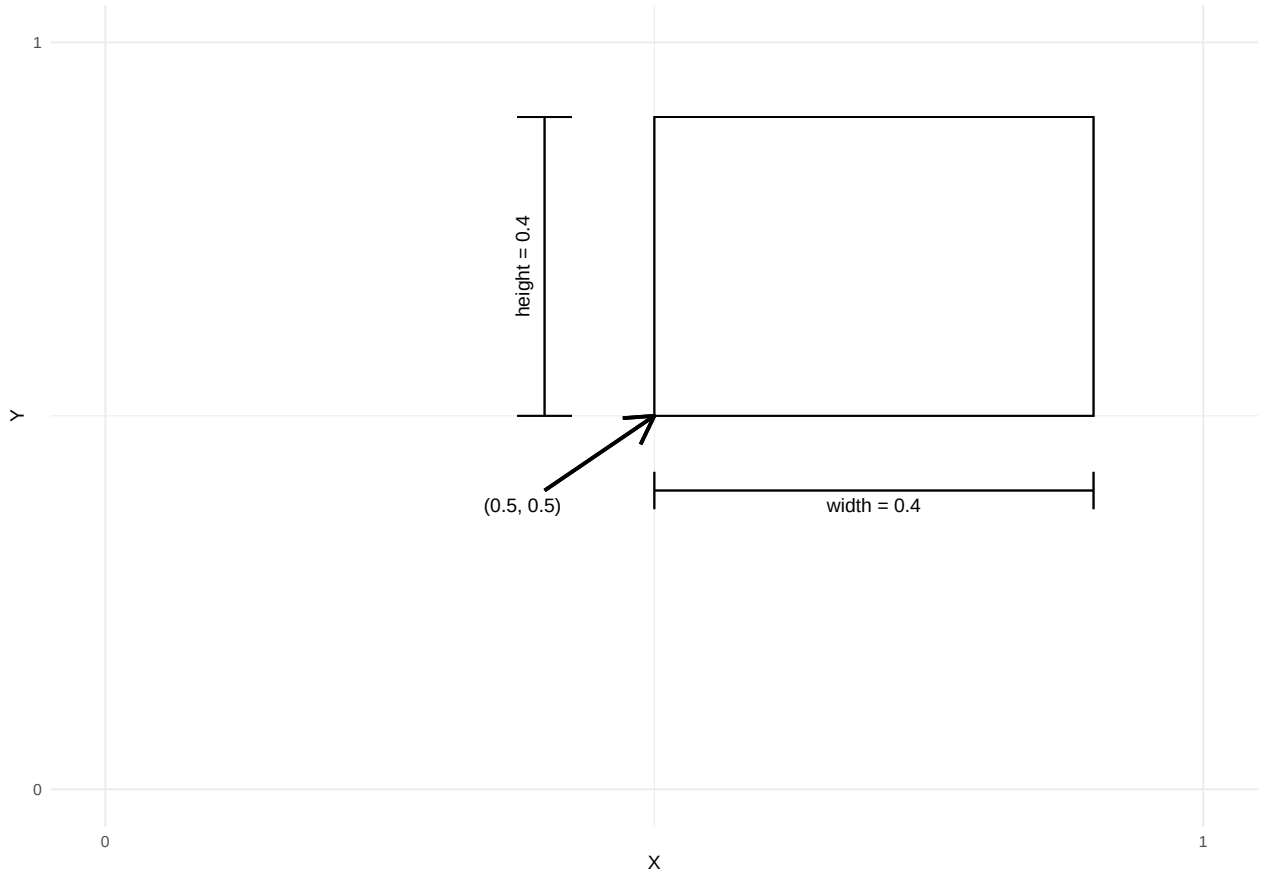
Patient characteristics by high tumour grade				
Characteristic	**Overall, N = 80[1]**	**No, N = 60[1]**	**Yes, N = 20[1]**	**p-value[2]**
Sex				0.6
Female	36 (45%)	26 (43%)	10 (50%)	
Male	44 (55%)	34 (57%)	10 (50%)	
Age (years)	61 (56, 65)	61 (56, 65)	60 (56, 64)	0.5
Adverse tumour marker	21 (26%)	9 (15%)	12 (60%)	**<0.001**
Tumour size (cm)	2.84 (1.97, 3.65)	2.34 (1.82, 3.17)	4.00 (3.45, 4.59)	**<0.001**
Diabetes	10 (12%)	7 (12%)	3 (15%)	0.7
Heart disease	29 (36%)	18 (30%)	11 (55%)	**0.044**
Hospital type				**<0.001**
Academic hospital	38 (48%)	22 (37%)	16 (80%)	
Community hospital	42 (52%)	38 (63%)	4 (20%)	
Extended surgical resection	36 (45%)	22 (37%)	14 (70%)	**0.009**
Postoperative complication	30 (38%)	23 (38%)	7 (35%)	0.8
Patient outcome				**0.016**
Alive	30 (38%)	27 (45%)	3 (15%)	
Died	50 (62%)	33 (55%)	17 (85%)	
Follow up (months)	39 (33, 50)	39 (34, 52)	38 (32, 46)	0.4

[1]n (%); Median (IQR)

[2]Pearson's Chi-squared test; Wilcoxon rank sum test; Fisher's exact test

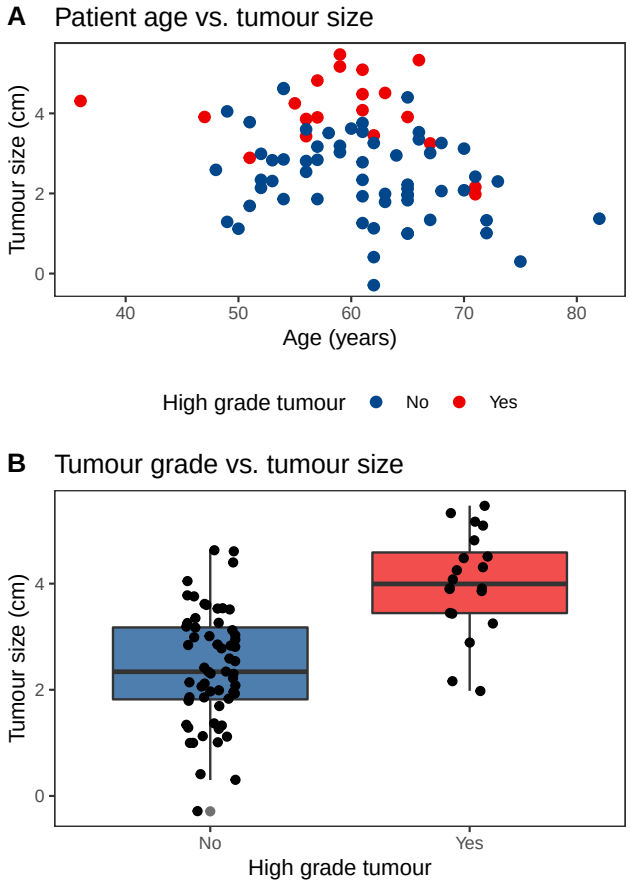
Cowplot

The plotting area is a 1x1 square and objects are placed according to the bottom left corner with defined heights and widths



Arrange figures with table using cowplot

```
ggdraw() +
  draw_plot(scatter_plot, x = 0, y = 0.5, width = 0.4, height = 0.4)+
  draw_plot(boxplot, x = 0, y = 0.1, width = 0.4, height = 0.4)+
  draw_plot(table_plot, x = 0.4, y = 0, width = 0.6, height = 1)+
  draw_plot_label(label = c('A', 'B', 'C'), size = 12,
                  x = c(0, 0, 0.4), y = c(0.9, 0.5, 0.9))
```



C

Patient characteristics by high tumour grade				
Characteristic	**Overall, N = 80[1]**	**No, N = 60[1]**	**Yes, N = 20[1]**	**p-value[2]**
Sex				0.6
Female	36 (45%)	26 (43%)	10 (50%)	
Male	44 (55%)	34 (57%)	10 (50%)	
Age (years)	61 (56, 65)	61 (56, 65)	60 (56, 64)	0.5
Adverse tumour marker	21 (26%)	9 (15%)	12 (60%)	**<0.001**
Tumour size (cm)	2.84 (1.97, 3.65)	2.34 (1.82, 3.17)	4.00 (3.45, 4.59)	**<0.001**
Diabetes	10 (12%)	7 (12%)	3 (15%)	0.7
Heart disease	29 (36%)	18 (30%)	11 (55%)	**0.044**
Hospital type				**<0.001**
Academic hospital	38 (48%)	22 (37%)	16 (80%)	
Community hospital	42 (52%)	38 (63%)	4 (20%)	
Extended surgical resection	36 (45%)	22 (37%)	14 (70%)	**0.009**
Postoperative complication	30 (38%)	23 (38%)	7 (35%)	0.8
Patient outcome				**0.016**
Alive	30 (38%)	27 (45%)	3 (15%)	
Died	50 (62%)	33 (55%)	17 (85%)	
Follow up (months)	39 (33, 50)	39 (34, 52)	38 (32, 46)	0.4

[1]n (%); Median (IQR)

[2]Pearson's Chi-squared test; Wilcoxon rank sum test; Fisher's exact test

Thank you!

Resources

1. R gallery of graphs
2. Hadley Wickham's book on ggplot2
3. Winston Chang's book on ggplot2

Questions?

matthew.castelo@medportal.ca