STAT 451 - Visualizing Data - Autumn 2025

Ariane Ducellier

10/07/2025

Tutorial Tidyverse part 1

Today, we are going to focus on basic Ggplot functions. For more on Ggplot, I recommend this book:

Applied data visualization with R and ggplot2 : Create useful, elaborate, and visually appealing plots Moulik, Tania 2018; Birmingham, UK : Packt Publishing Ltd.

We will need the following R libraries:

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggpubr)
library(gridExtra)
```

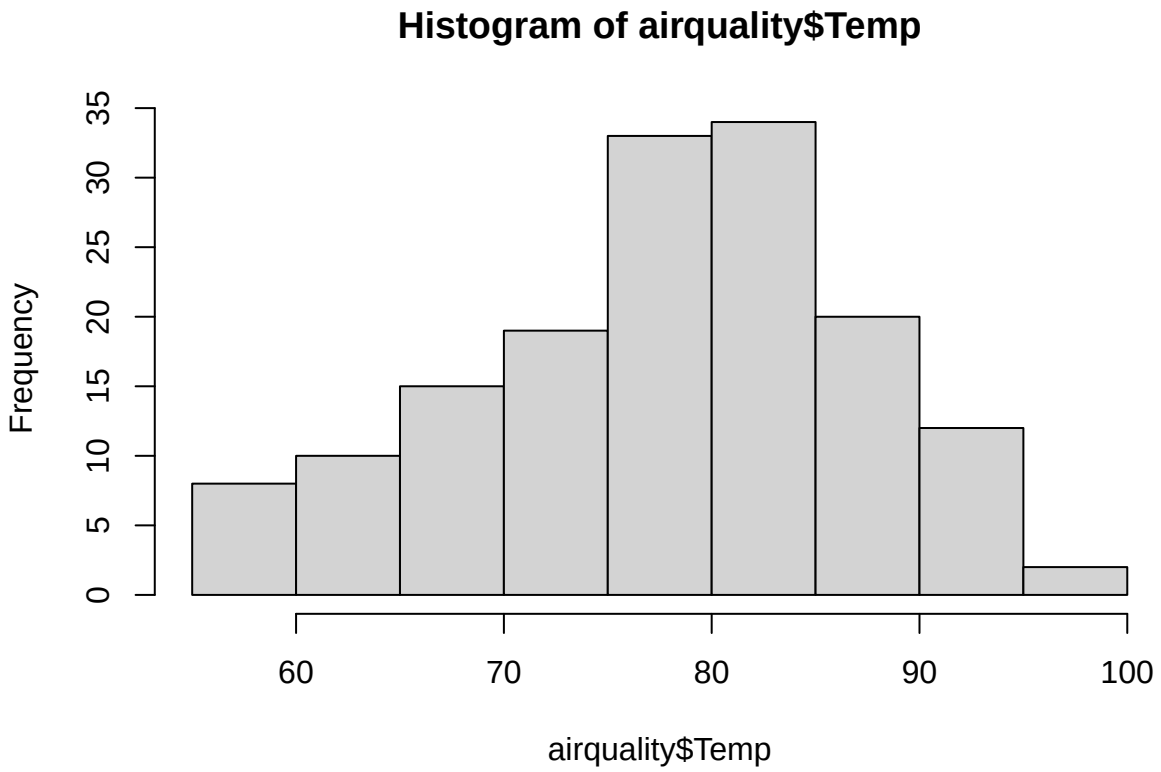
```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(Lock5Data)
```

Basic Plotting in ggplot2

Histograms

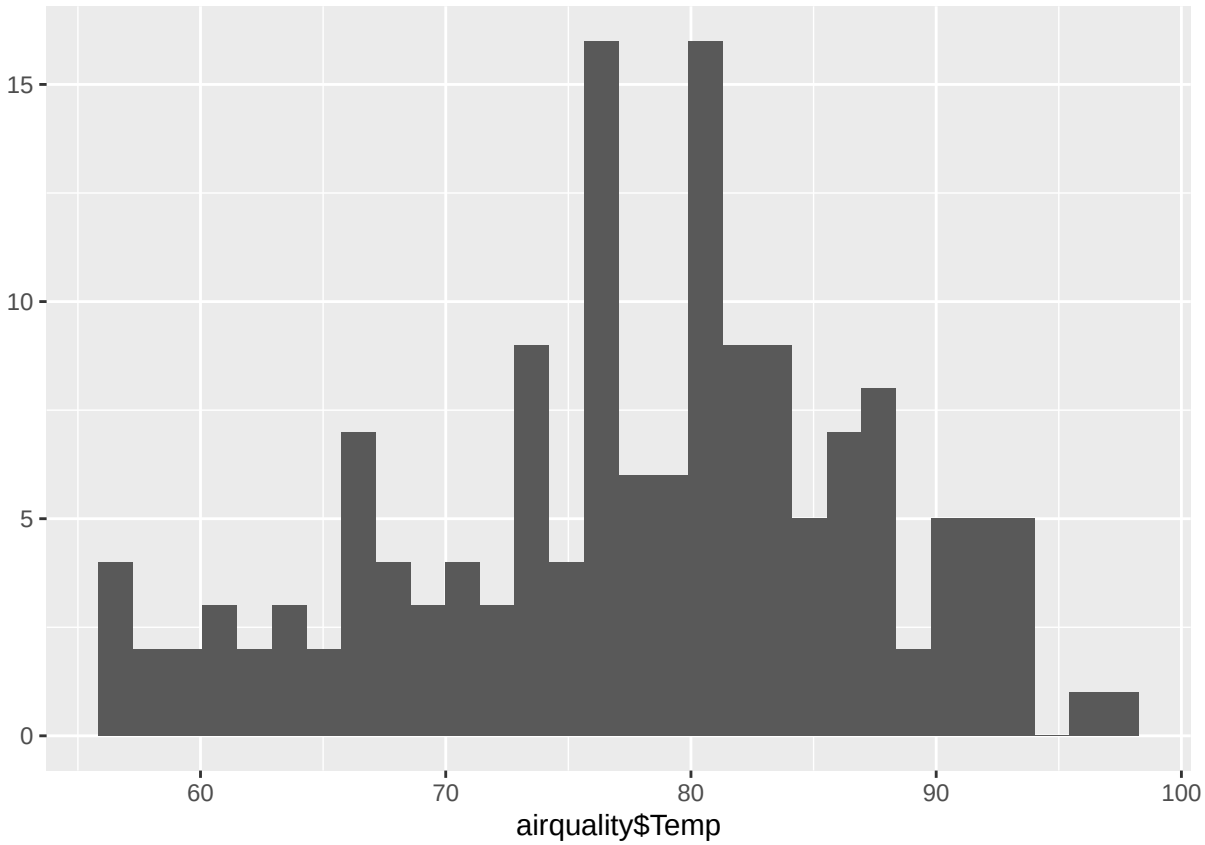
```
hist(airquality$Temp)
```




```
qplot(airquality$Temp)
```

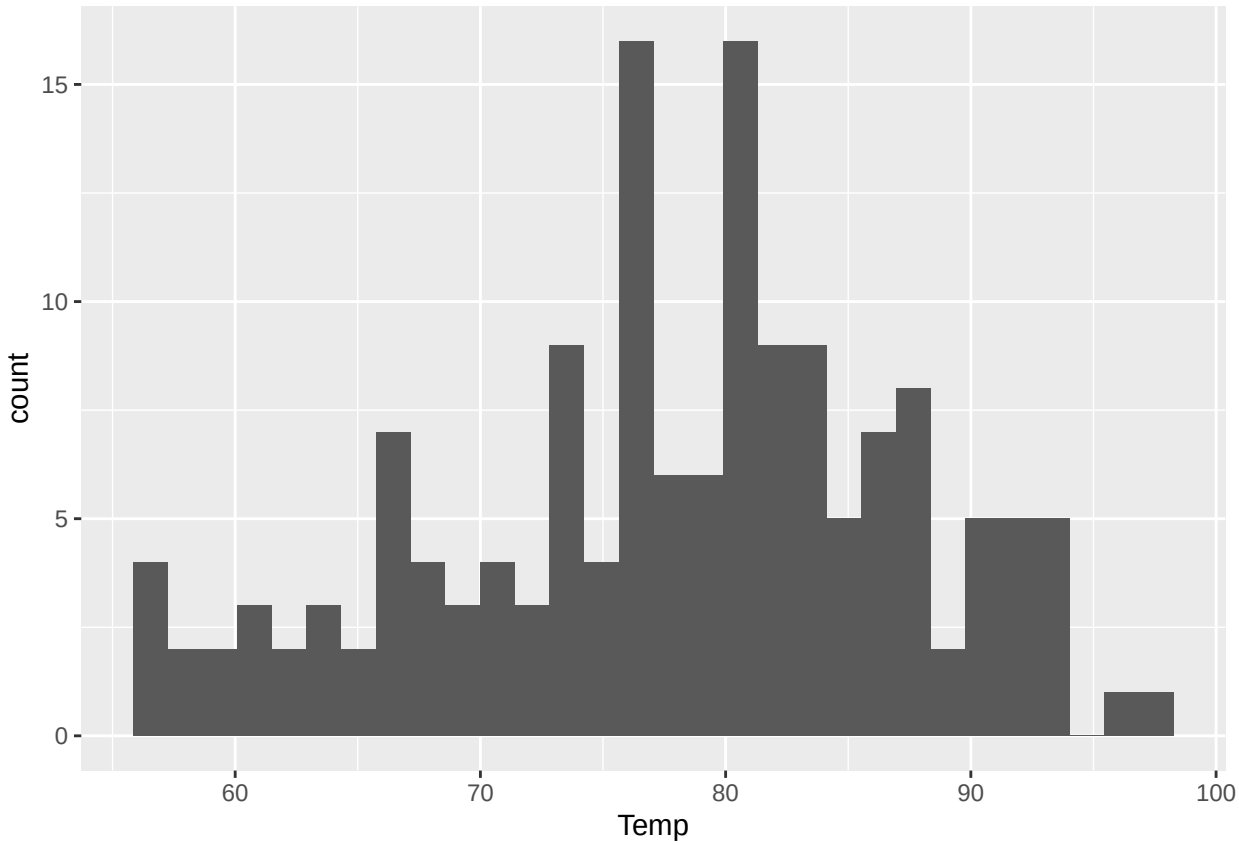
```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(airquality, aes(x=Temp)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

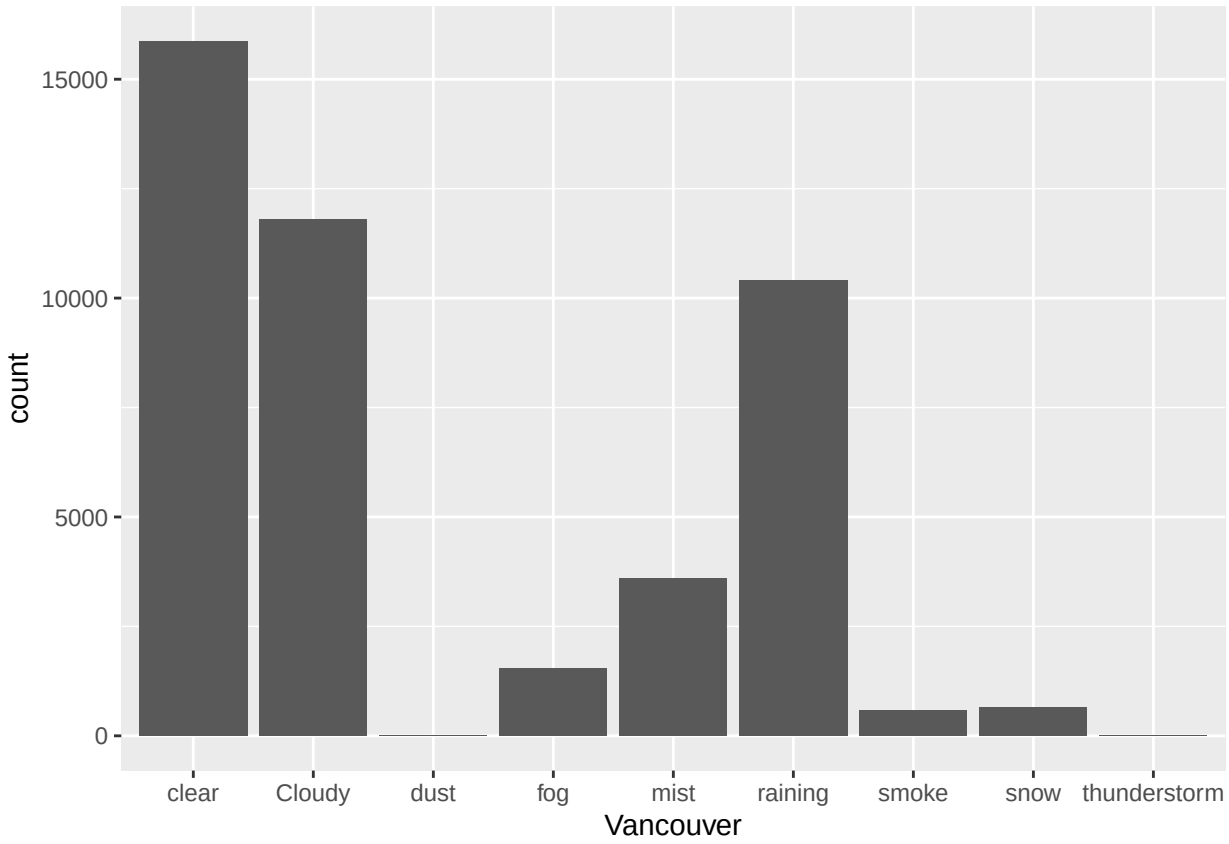


Bar plots

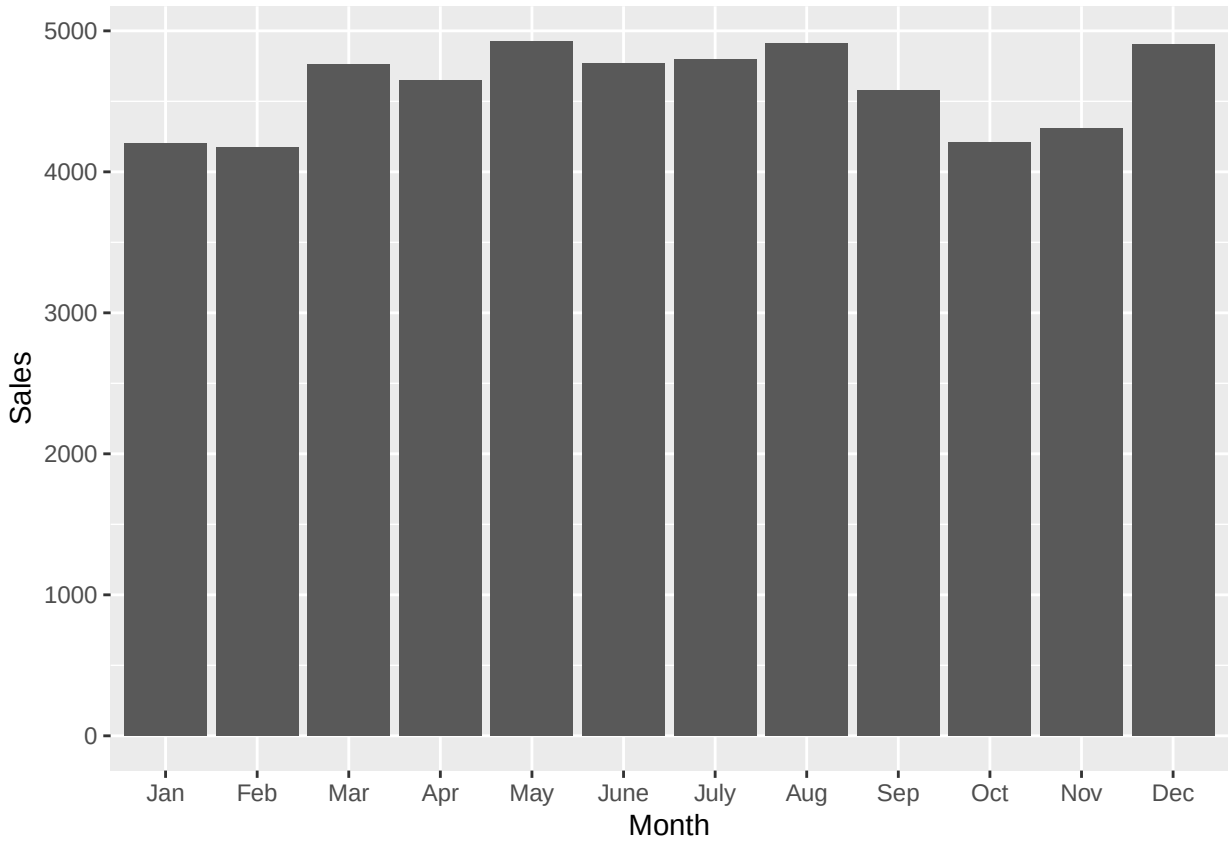
```
df_desc <- read_csv("../../data/historical-hourly-weather-data/weather_description.csv")
```

```
## New names:
## Rows: 44459 Columns: 4
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (3): Vancouver, Seattle, San.Francisco dbl (1): ...1
## 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.
## * `` -> `...1`
```

```
ggplot(df_desc, aes(x=Vancouver)) + geom_bar()
```




```
df <- na.omit(RetailSales)
months_of_the_year <- c("Jan", "Feb", "Mar", "Apr", "May", "June",
                        "July", "Aug", "Sep", "Oct", "Nov", "Dec")
ggplot(df) +
geom_bar(aes(x=factor(Month, months_of_the_year), y=Sales), stat="identity") +
xlab("Month")
```



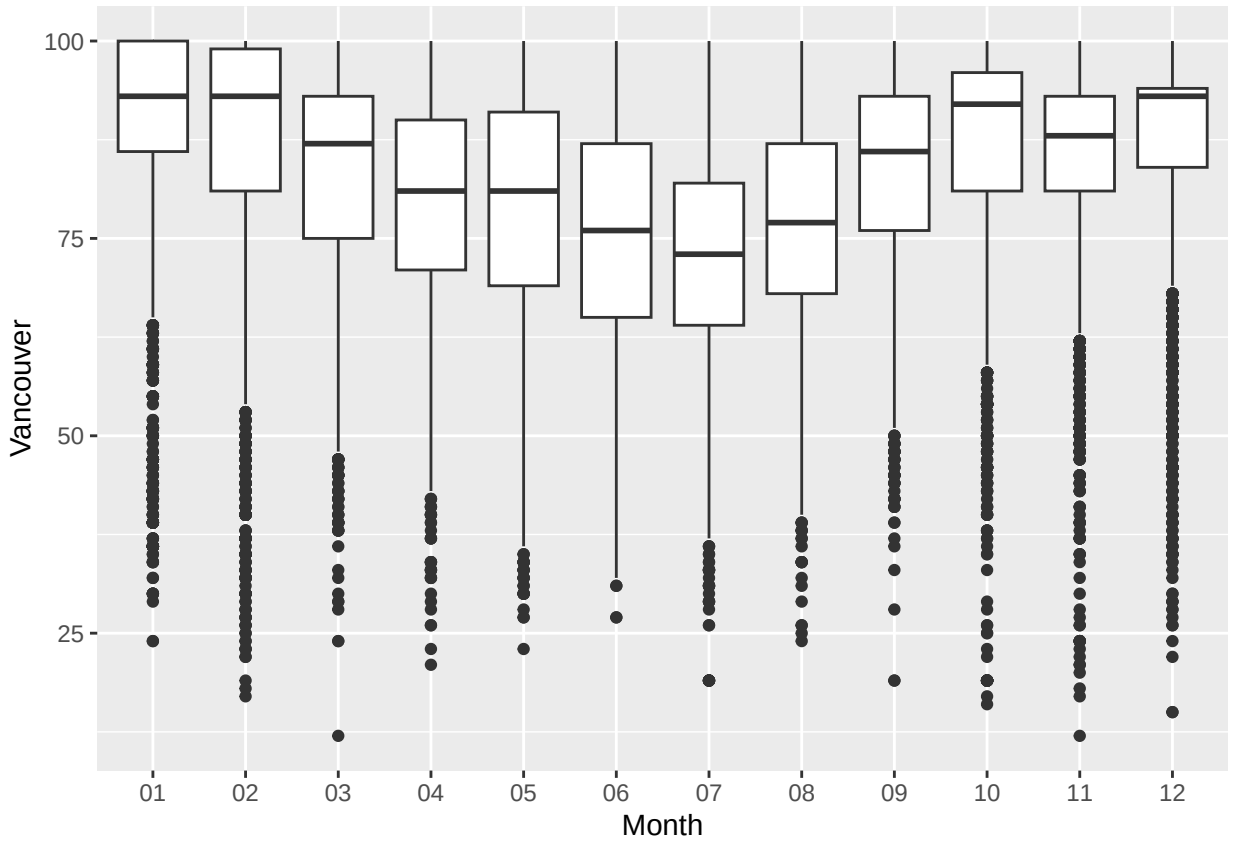
Box plots

```
df_hum <- read_csv("../../data/historical-hourly-weather-data/humidity.csv")
```

```
## Rows: 45253 Columns: 37
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl  (36): Vancouver, Portland, San Francisco, Seattle, Los Angeles, San Die...
## dttm  (1): datetime
##
## 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.
```

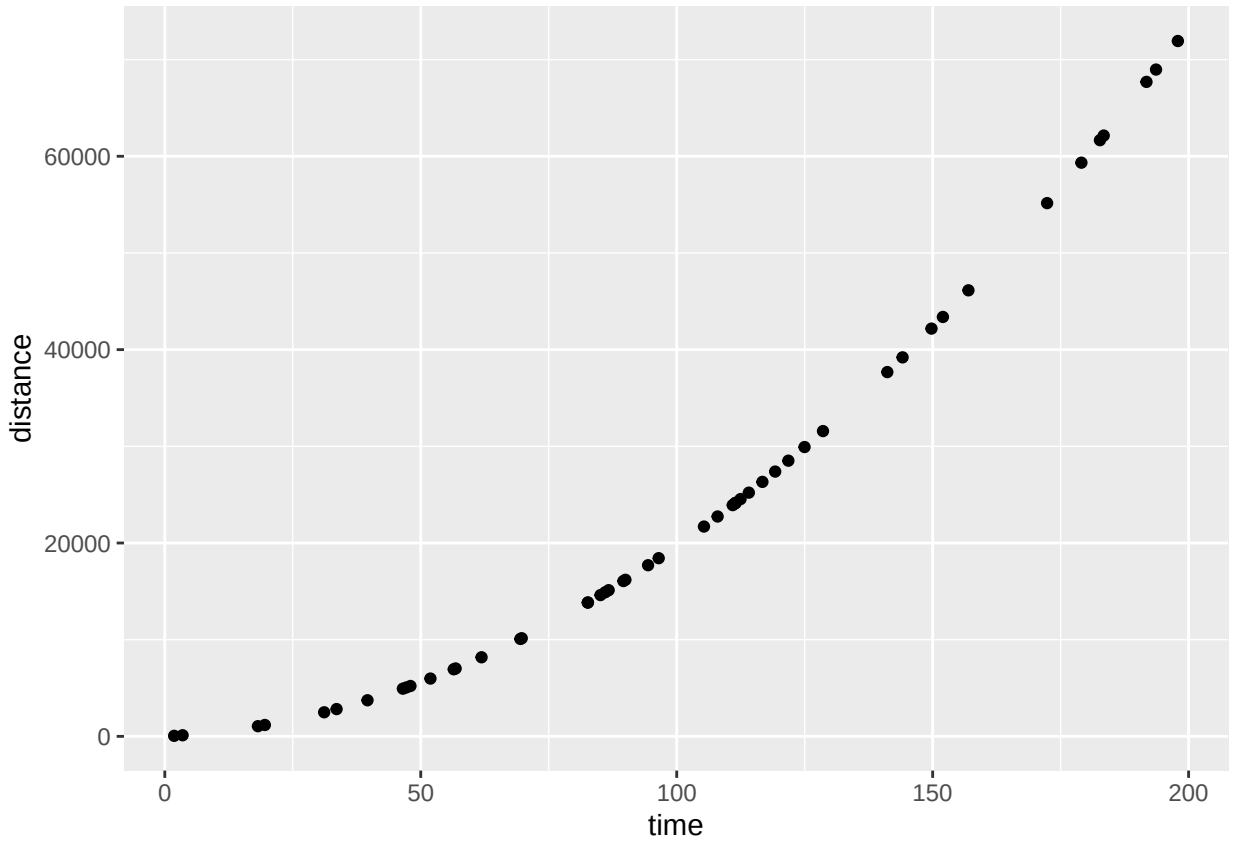
```
df_hum$datetime <- as.character(df_hum$datetime)
df_hum$Month <- substr(df_hum$datetime, 6, 7)
ggplot(df_hum, aes(x=Month, y=Vancouver)) +
geom_boxplot()
```

```
## Warning: Removed 1826 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

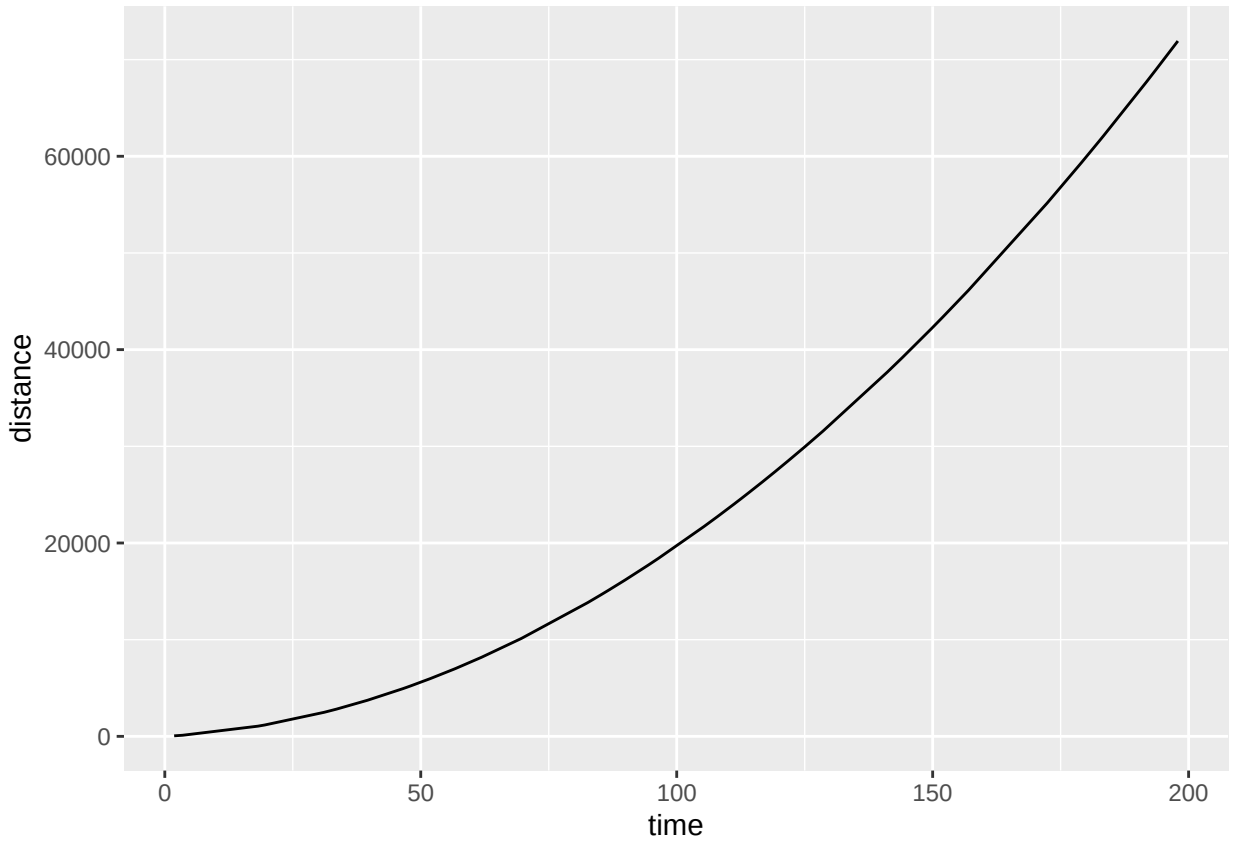


Scatter plots and line plots

```
a = 3.4
v0 = 27
time <- runif(50, min=0, max=200)
distance <- sapply(time, function(x) v0 * x + 0.5 * a * x^2)
df <- data.frame(time,distance)
ggplot(df, aes(x=time, y=distance)) + geom_point()
```




```
ggplot(df, aes(x=time, y=distance)) + geom_line()
```



Changing histogram defaults and adding aesthetics

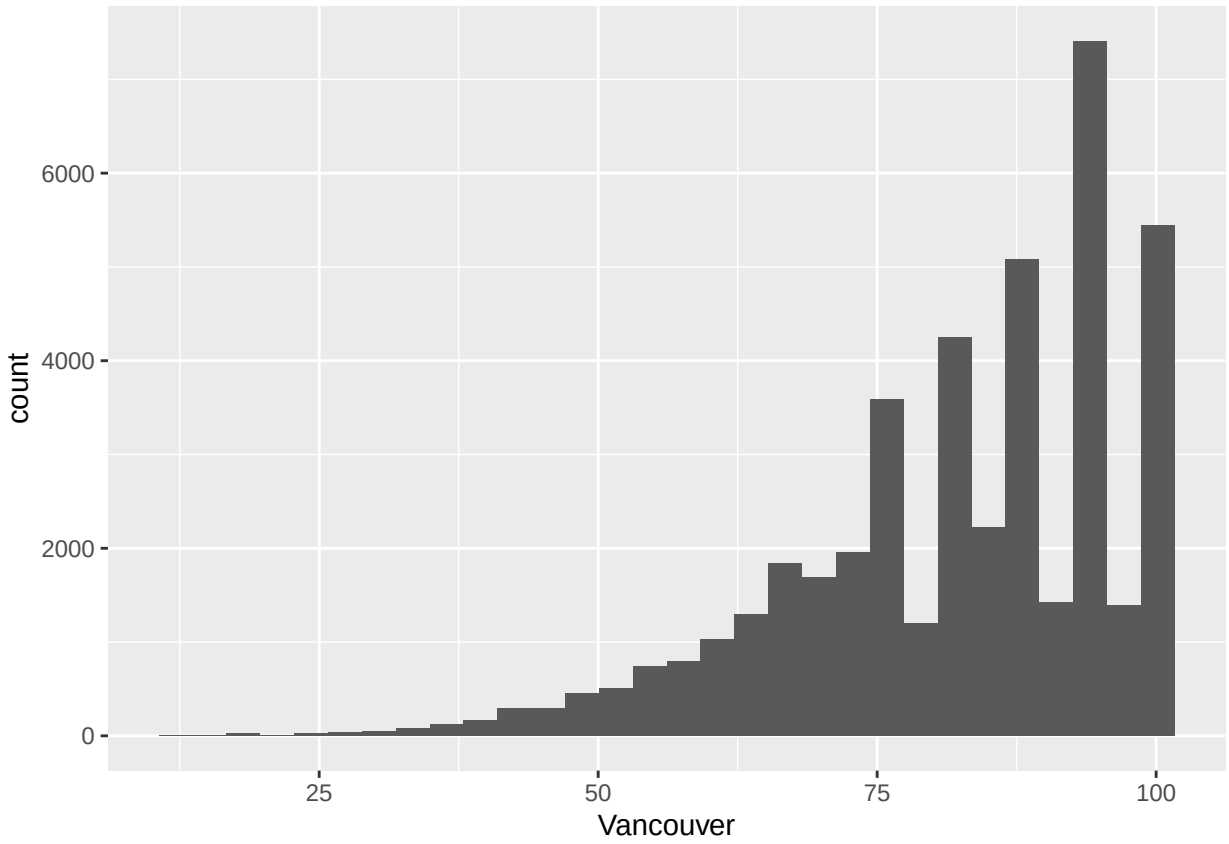
```
df_hum <- read_csv("../../data/historical-hourly-weather-data/humidity.csv")
```

```
## Rows: 45253 Columns: 37
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl  (36): Vancouver, Portland, San Francisco, Seattle, Los Angeles, San Die...
## dttm  (1): datetime
##
## 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.
```

```
ggplot(df_hum, aes(x=Vancouver)) + geom_histogram()
```

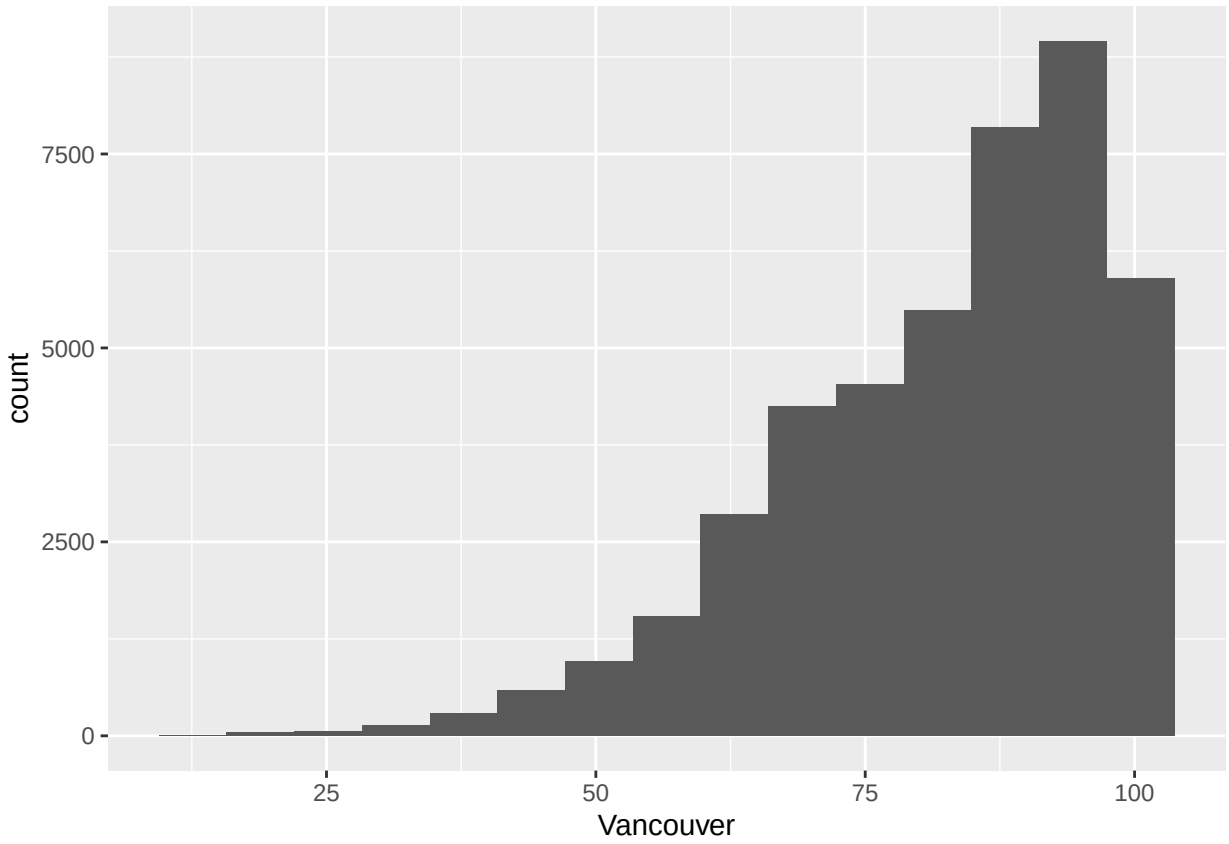
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 1826 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



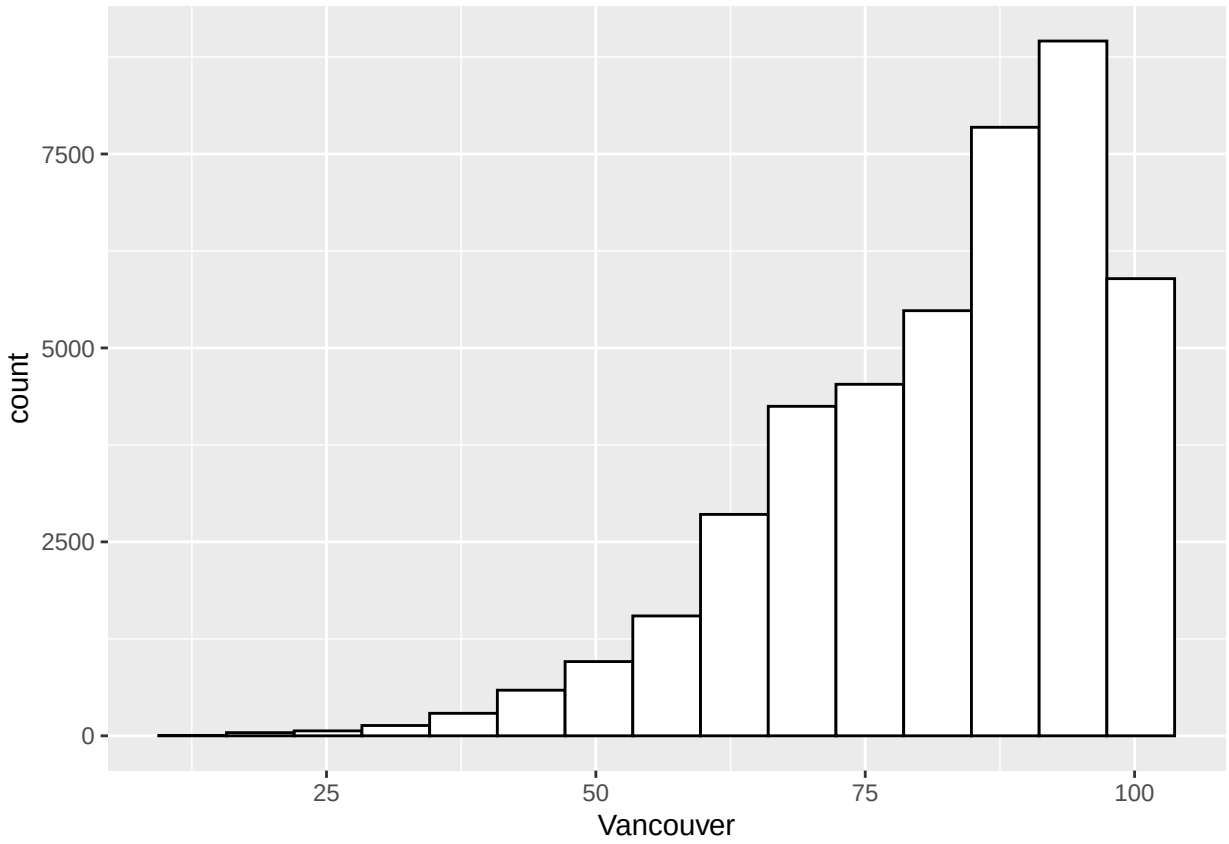

```
ggplot(df_hum, aes(x=Vancouver)) + geom_histogram(bins=15)
```

```
## Warning: Removed 1826 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



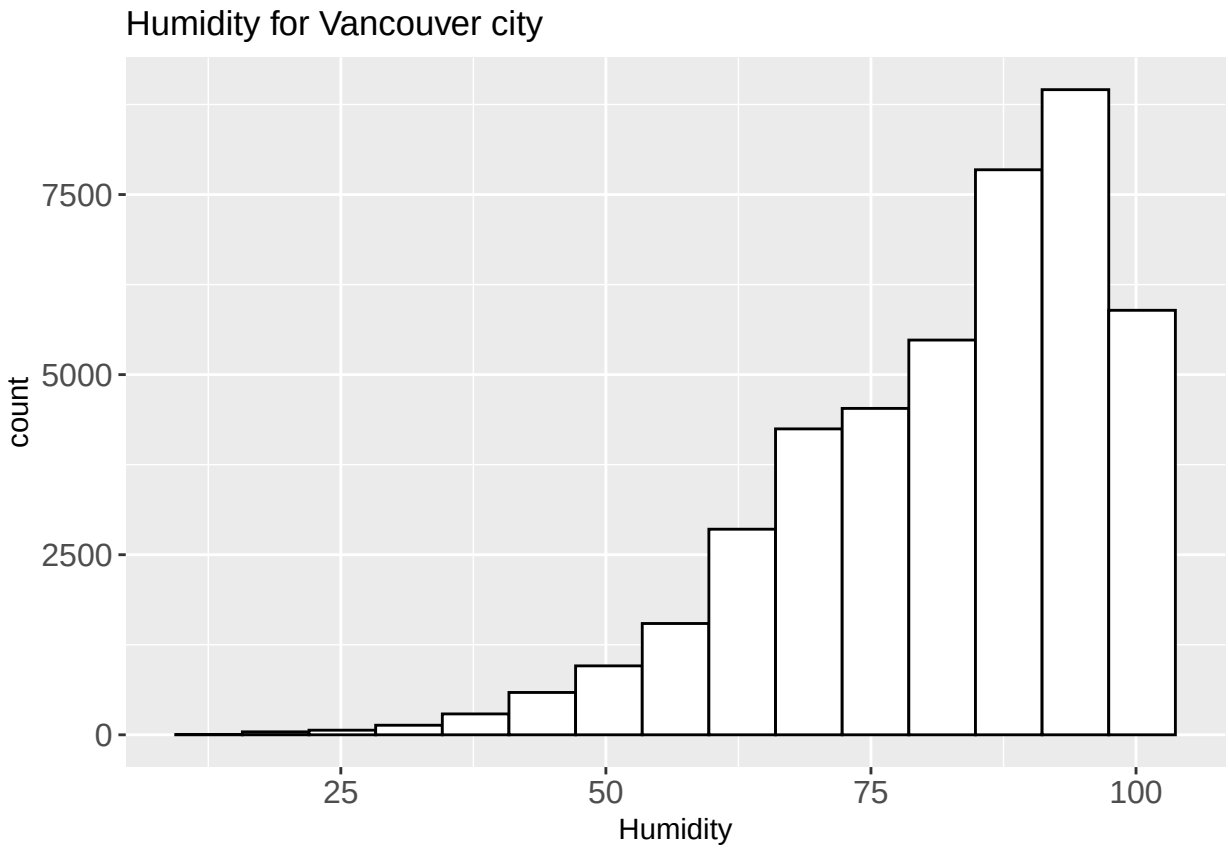

```
ggplot(df_hum, aes(x=Vancouver)) + geom_histogram(bins=15, fill="white", color=1)
```

```
## Warning: Removed 1826 rows containing non-finite outside the scale range
## (`stat_bin()`).
```




```
ggplot(df_hum, aes(x=Vancouver)) +
geom_histogram(bins=15, fill="white", color=1) +
ggtitle("Humidity for Vancouver city") +
xlab("Humidity") +
theme(axis.text.x=element_text(size=12), axis.text.y=element_text(size=12))
```

```
## Warning: Removed 1826 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



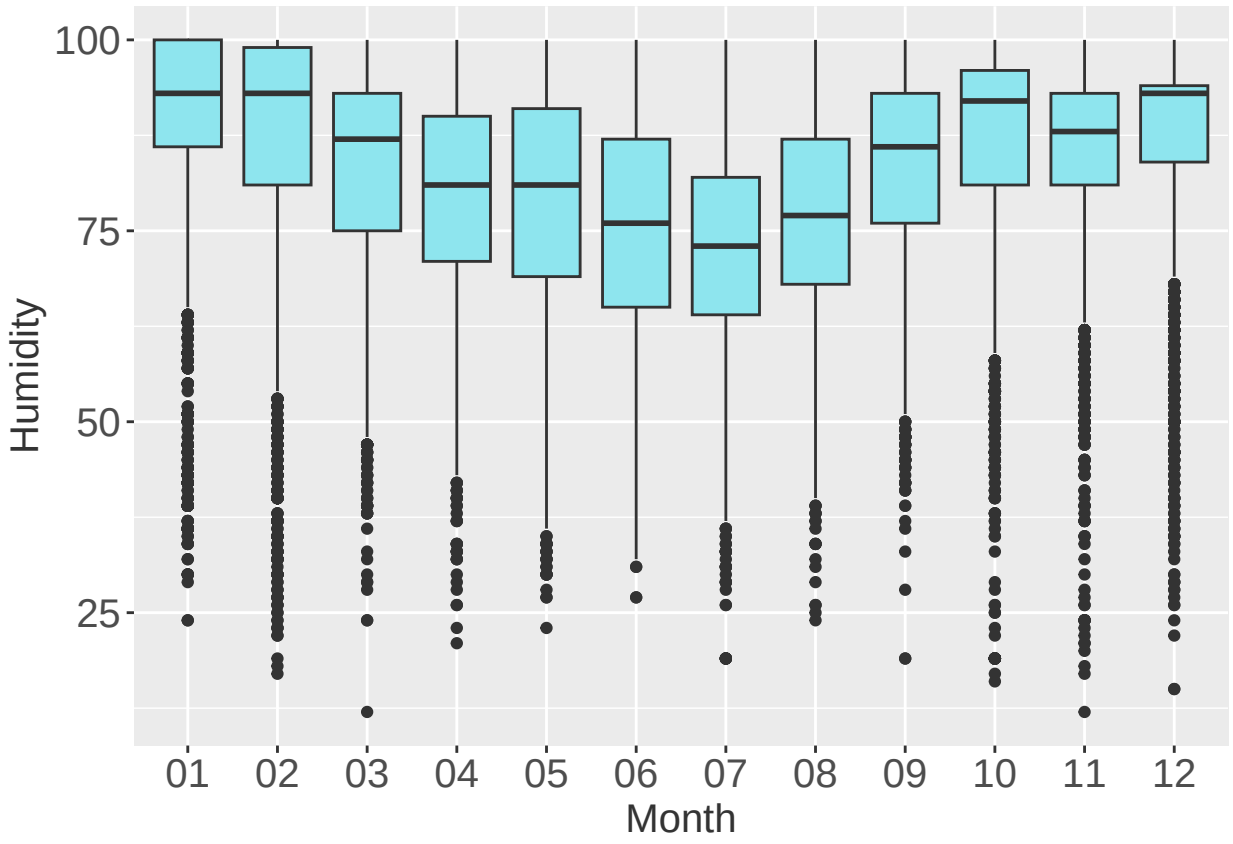
Changing boxplot defaults and adding aesthetics

```
df_hum <- read_csv("../../data/historical-hourly-weather-data/humidity.csv")
```

```
## Rows: 45253 Columns: 37
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl  (36): Vancouver, Portland, San Francisco, Seattle, Los Angeles, San Die...
## dttm  (1): datetime
##
## 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.
```

```
df_hum$datetime <- as.character(df_hum$datetime)
df_hum$Month <- substr(df_hum$datetime, 6, 7)
ggplot(df_hum, aes(x=Month, y=Vancouver)) +
geom_boxplot(color="gray20", fill="cadetblue2") +
ylab("Humidity") +
theme(axis.text.x=element_text(size=15),
axis.text.y=element_text(size=15),
axis.title.x=element_text(size=15, color="gray20"),
axis.title.y=element_text(size=15, color="gray20"))
```

```
## Warning: Removed 1826 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



Grammar of Graphics and Visual Components

Layers

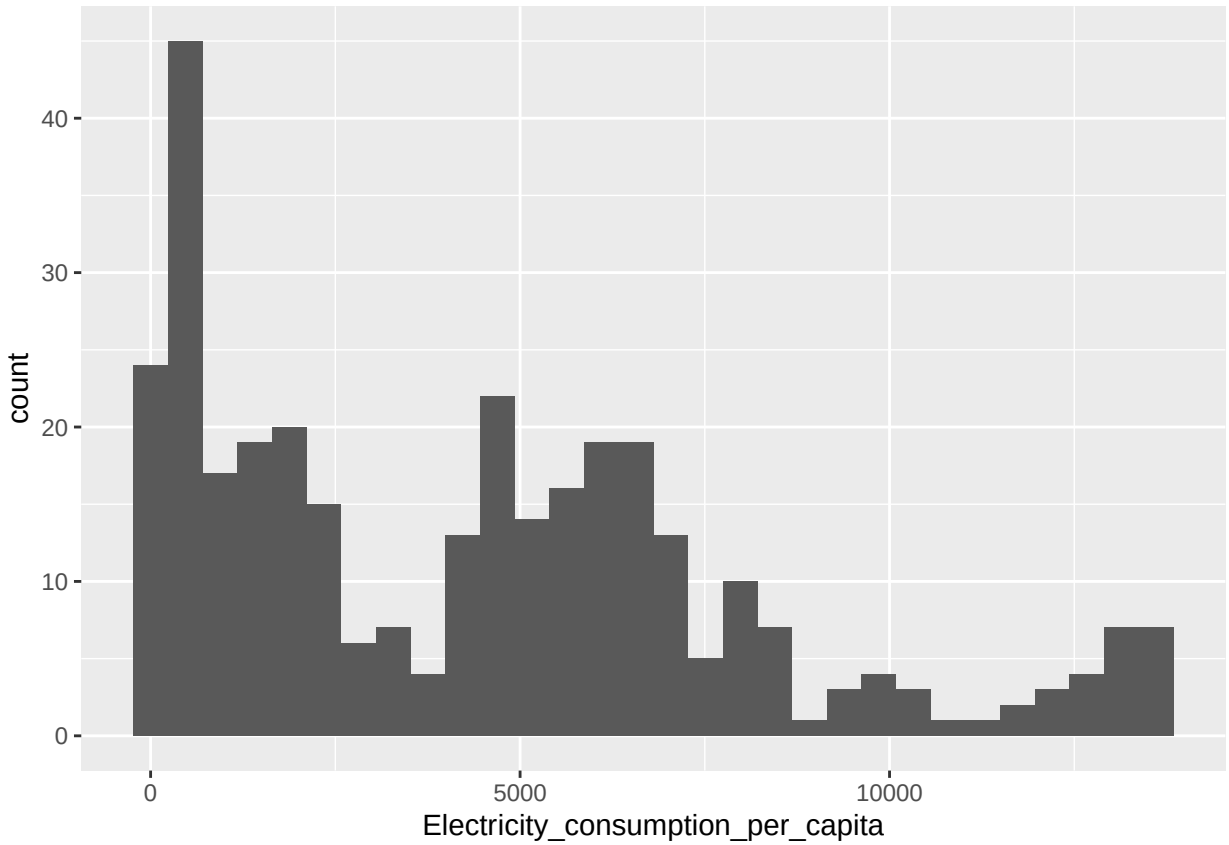
```
df <- read_csv("../../data/gapminder-data.csv")
```

```
## New names:
## Rows: 1512 Columns: 10
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (1): Country dbl (9): ...1, Year, gdp_per_capita,
## Electricity_consumption_per_capita, und...
## 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.
## * `` -> `...1`
```

```
p1 <- ggplot(df, aes(x=Electricity_consumption_per_capita))
p2 <- p1 + geom_histogram()
p2
```

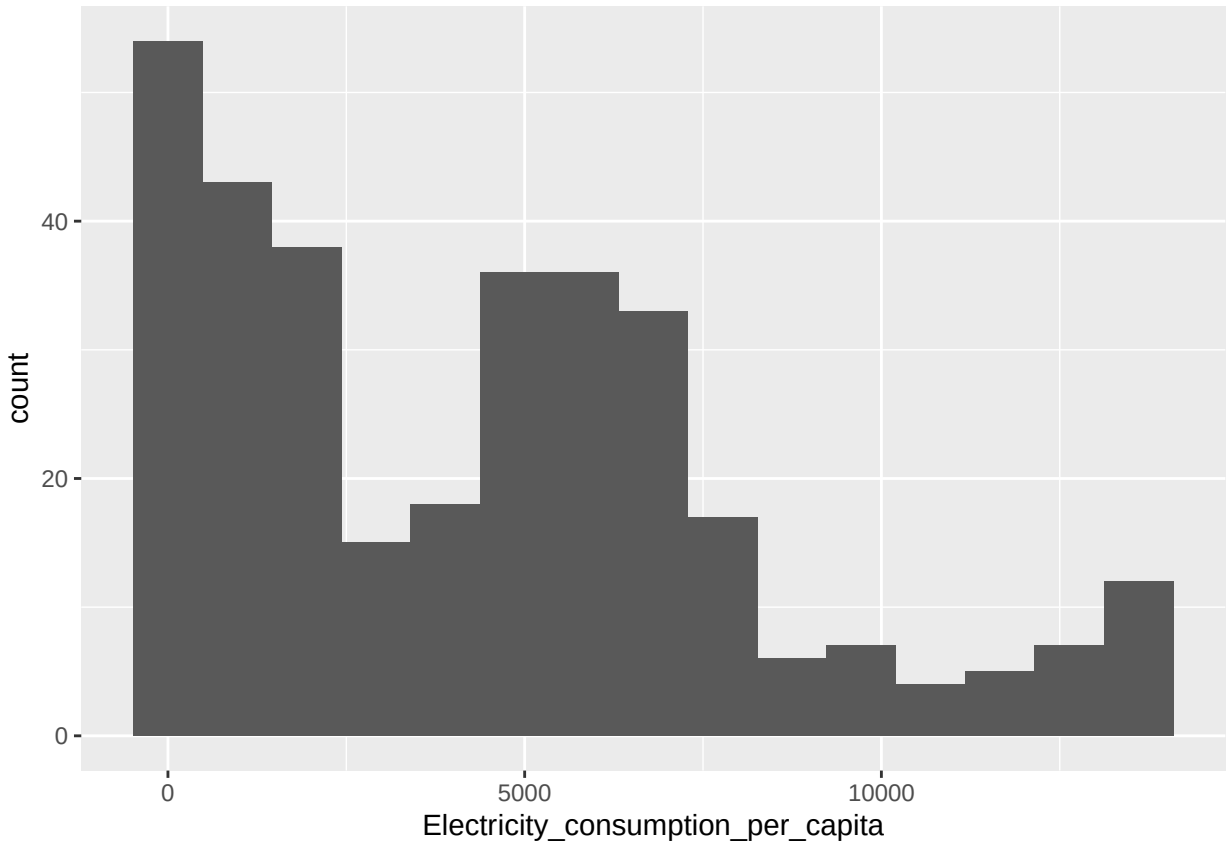
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 1181 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



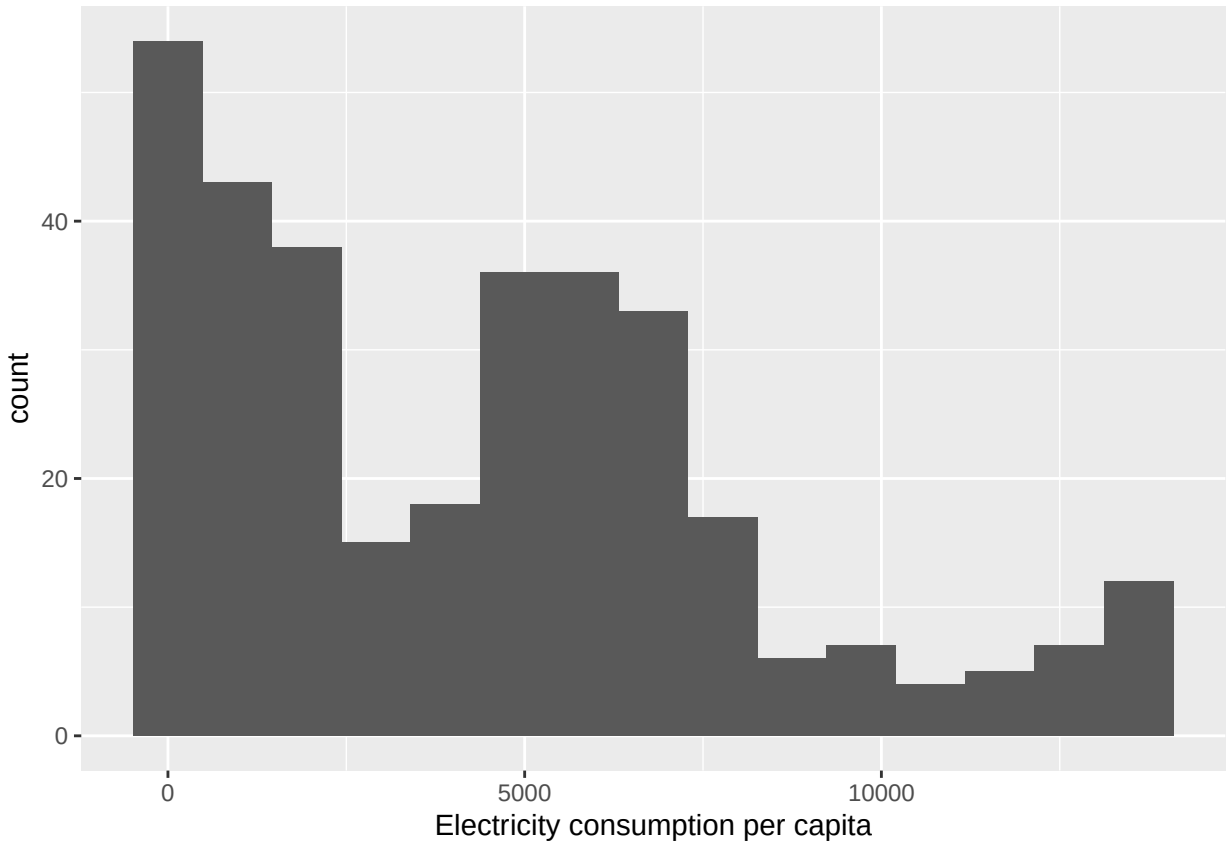

```
p3 <- p1 + geom_histogram(bins=15)
p3
```

```
## Warning: Removed 1181 rows containing non-finite outside the scale range
## (`stat_bin()`).
```




```
p4 <- p3 + xlab("Electricity consumption per capita")
p4
```

```
## Warning: Removed 1181 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



Scales

```
df <- read_csv("../../data/gapminder-data.csv")
```

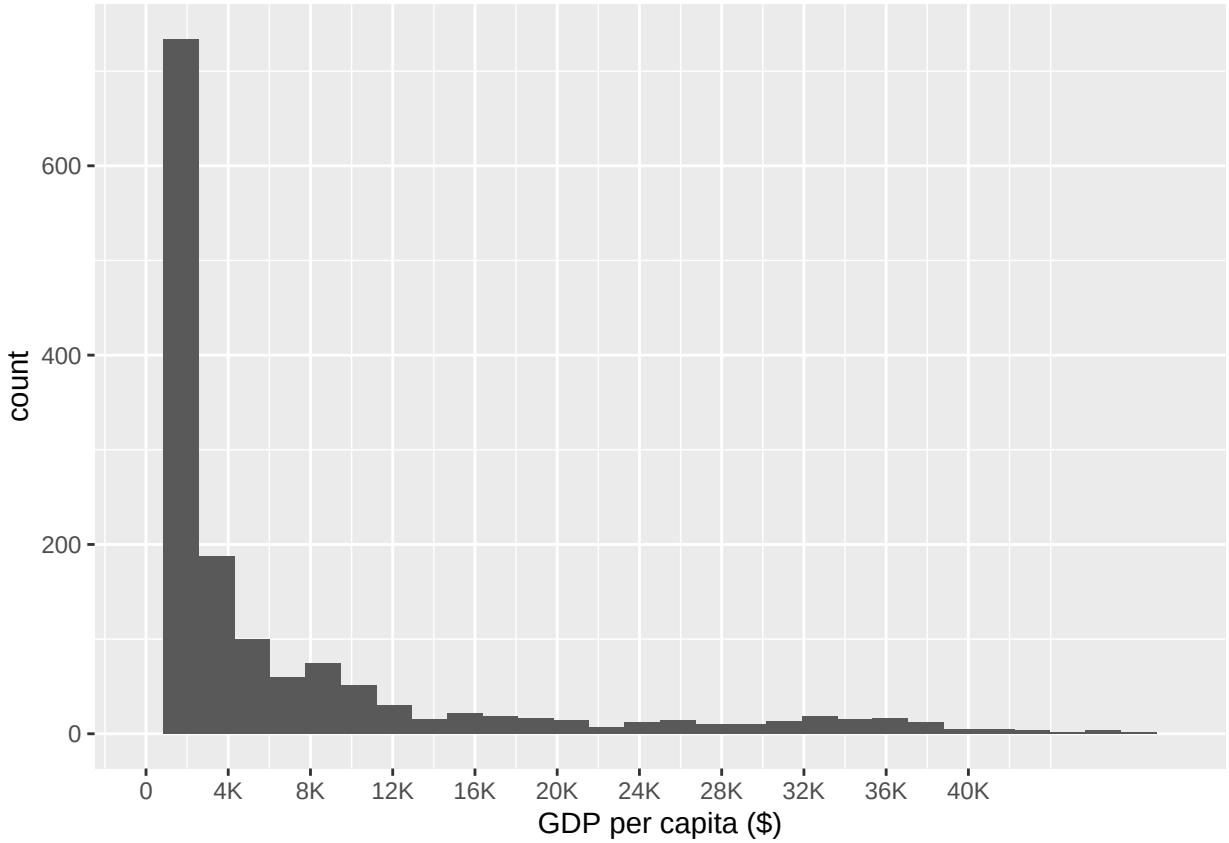
```
## New names:
## Rows: 1512 Columns: 10
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (1): Country dbl (9): ...1, Year, gdp_per_capita,
## Electricity_consumption_per_capita, und...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
p1 <- ggplot(df, aes(x=gdp_per_capita))
p2 <- p1 + geom_histogram()
p3 <- p2 + scale_x_continuous(name='GDP per capita ($)',
                              limits=c(0, 50000),
                              breaks=seq(0, 40000, 4000),
                              labels=c('0', '4K', '8K', '12K', '16K', '20K',
                                       '24K', '28K', '32K', '36K', '40K'))
p3
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

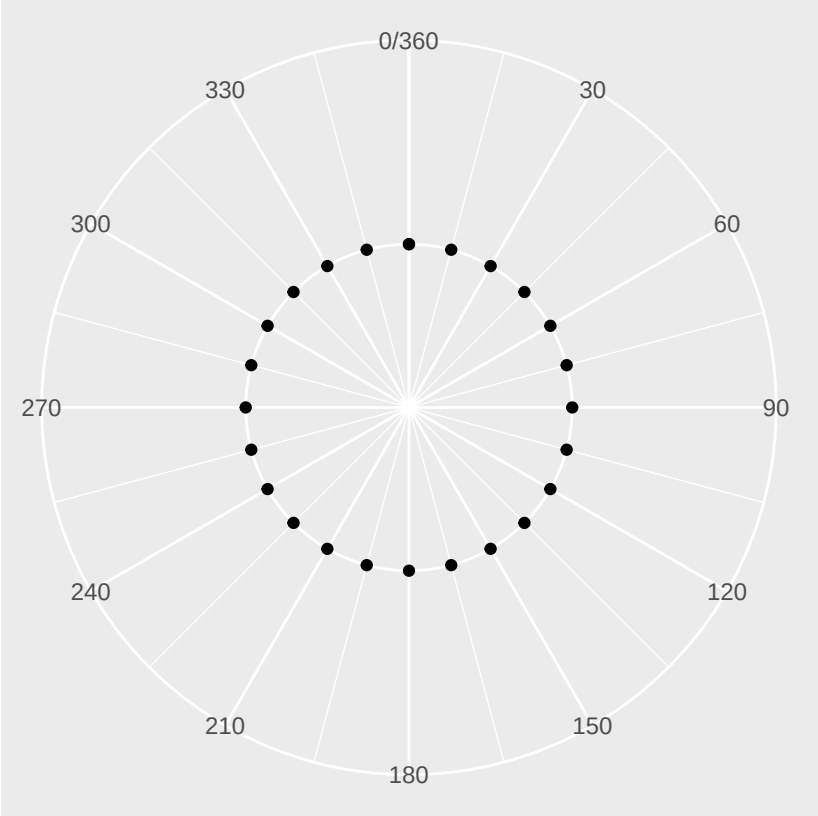
```
## Warning: Removed 7 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```



Polar coordinates

```
t <- seq(0, 360, by=15)
r <- 2
qplot(r, t) +
coord_polar(theta="y") +
scale_y_continuous(breaks=seq(0, 360, 30)) +
theme(axis.title.x=element_blank(),
      axis.title.y=element_blank(),
      axis.text.y=element_blank(),
      axis.ticks.y=element_blank())
```



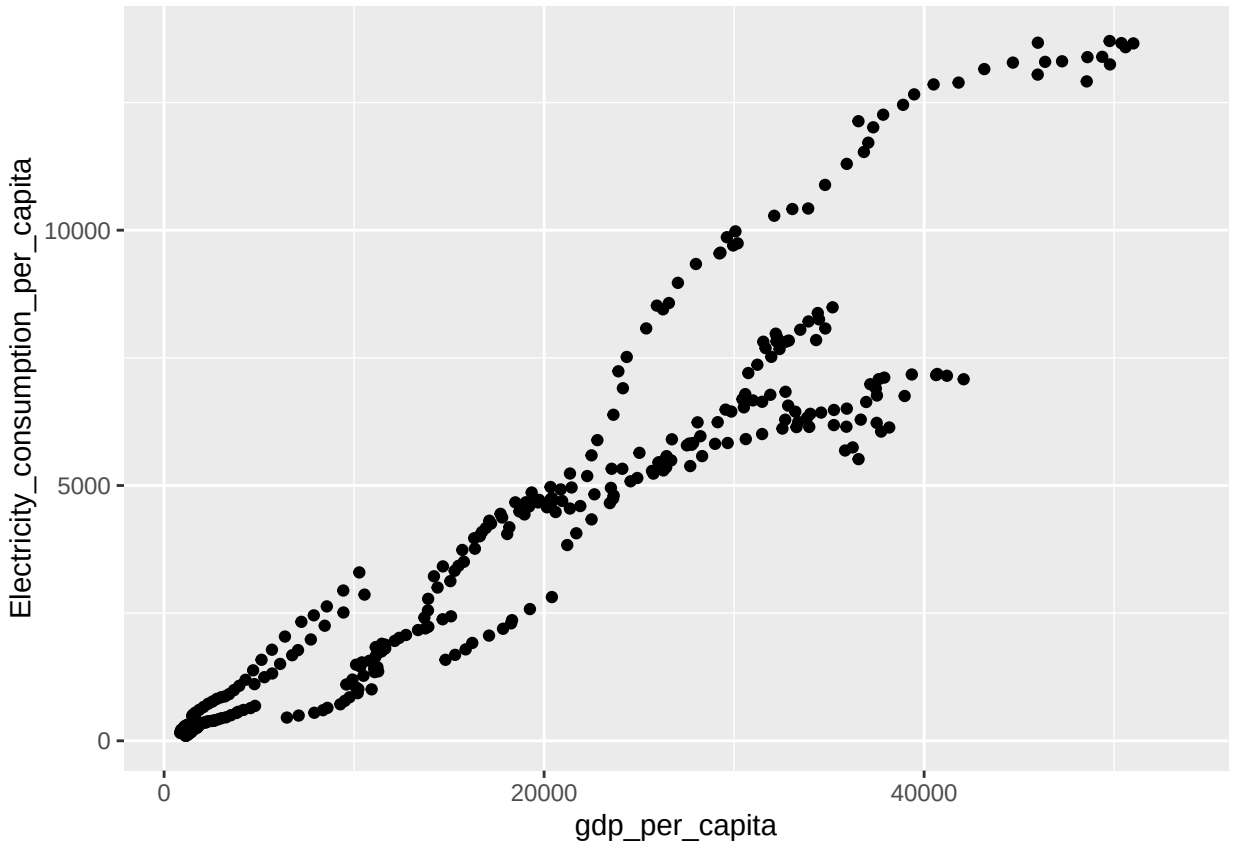
Facets

```
df <- read_csv("../../data/gapminder-data.csv")
```

```
## New names:
## Rows: 1512 Columns: 10
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (1): Country dbl (9): ...1, Year, gdp_per_capita,
## Electricity_consumption_per_capita, und...
## 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.
## * `` -> `...1`
```

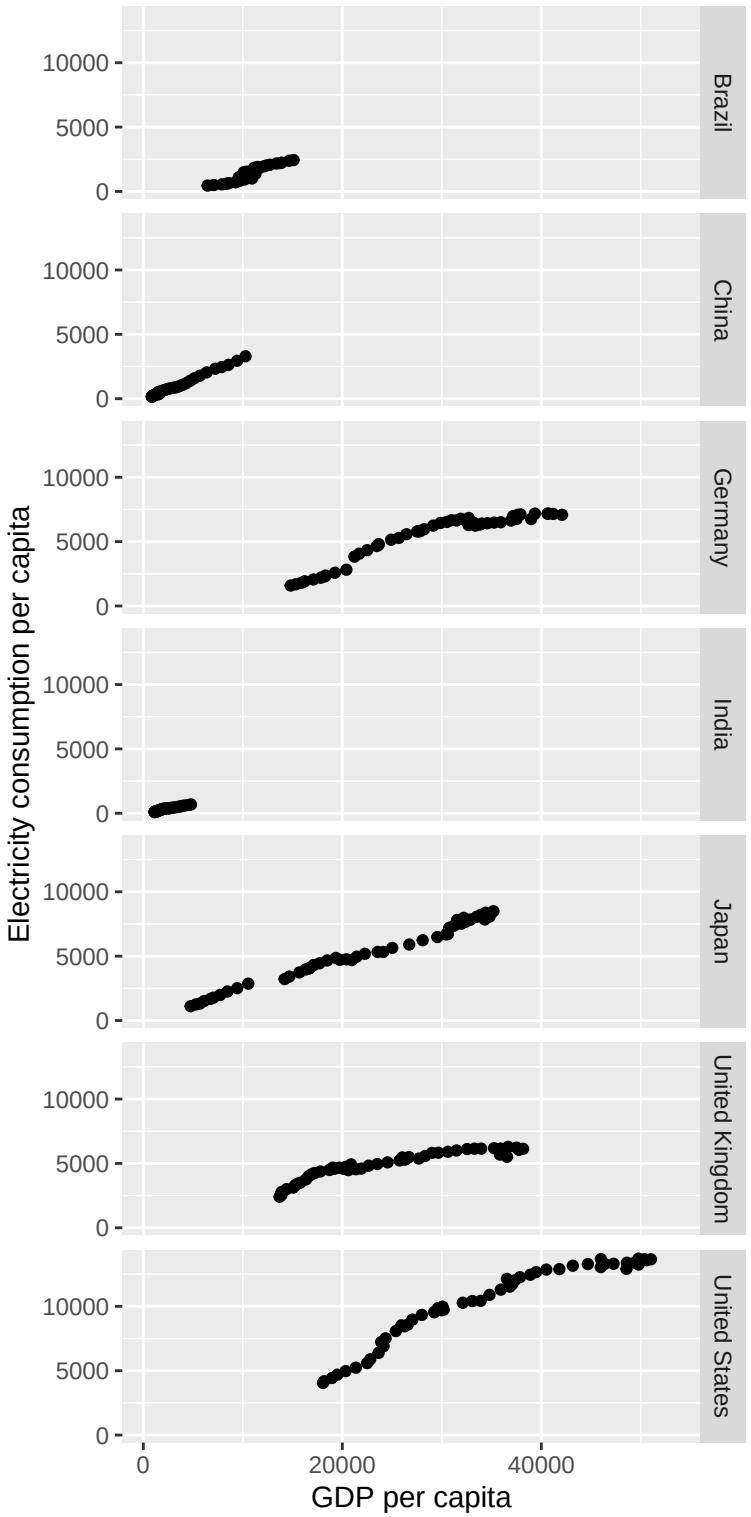
```
p <- ggplot(df, aes(x=gdp_per_capita, y=Electricity_consumption_per_capita)) + geom_point()
p
```

```
## Warning: Removed 1181 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
p + facet_grid(Country ~ .) +
xlab("GDP per capita") +
ylab("Electricity consumption per capita")
```

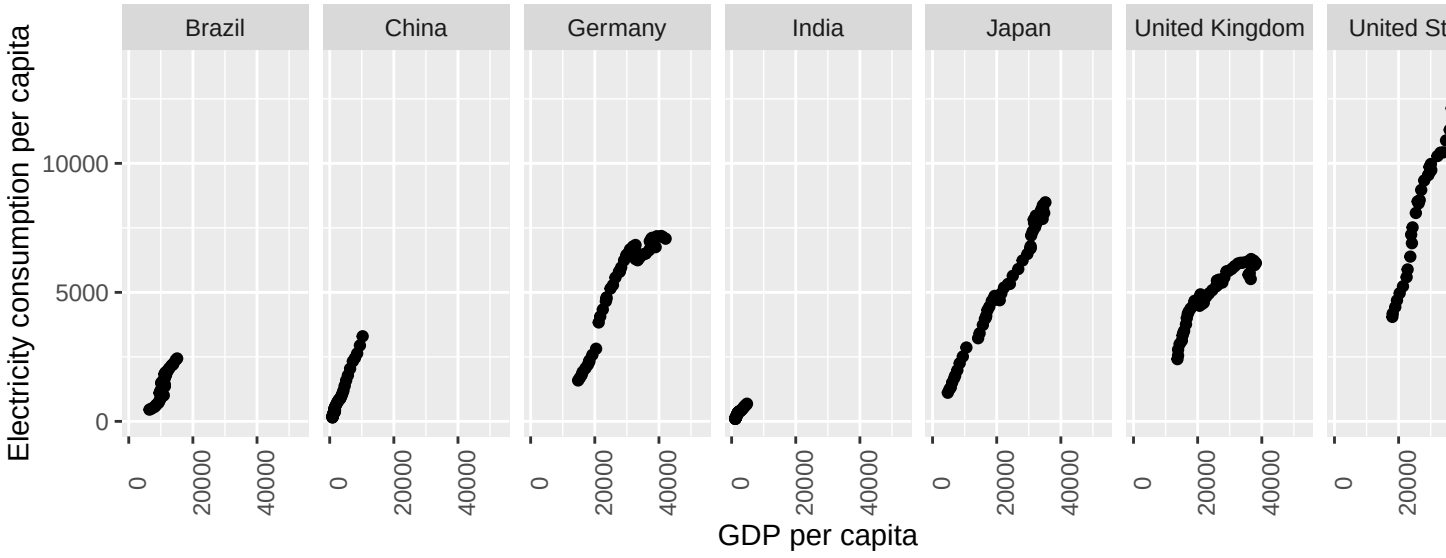
```
## Warning: Removed 1181 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
p + facet_grid(. ~ Country) +
xlab("GDP per capita") +
ylab("Electricity consumption per capita") +
theme(axis.text.x=element_text(angle=90))
```

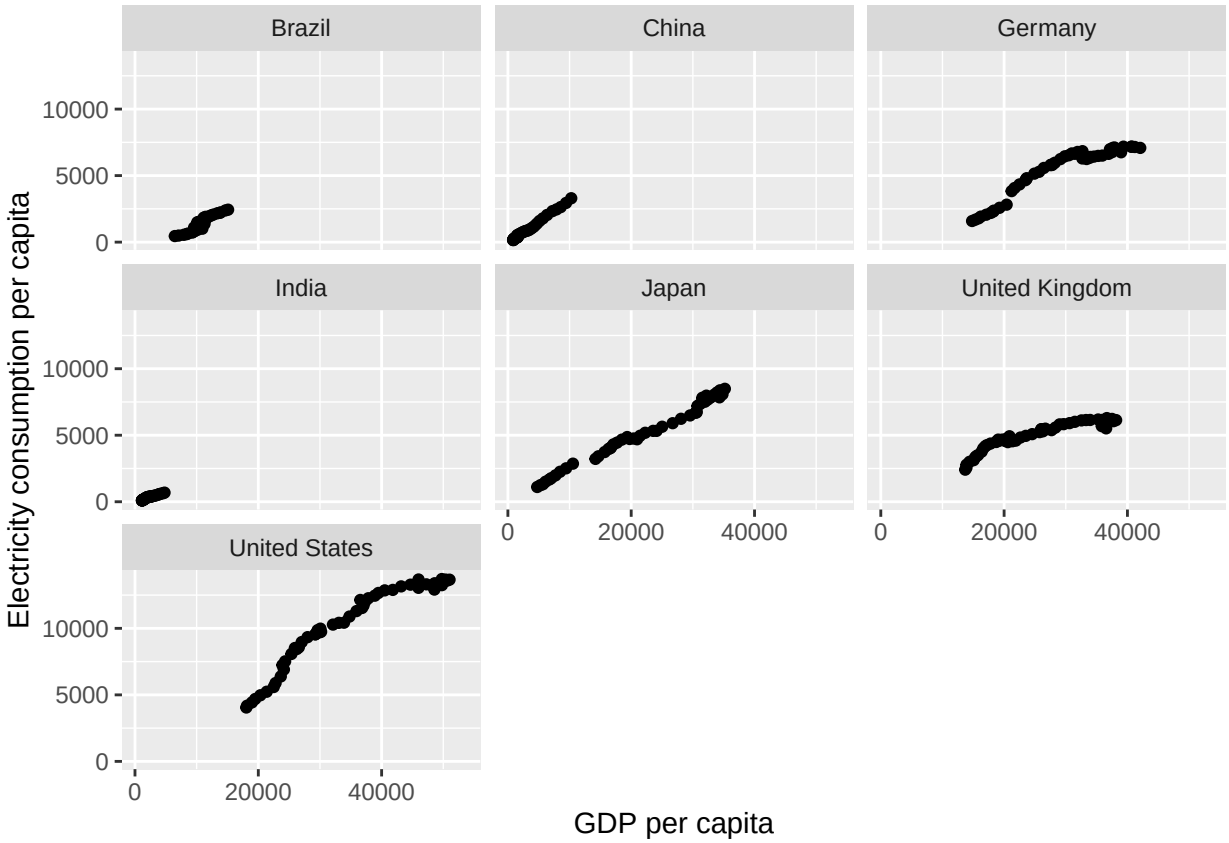
```
## Warning: Removed 1181 rows containing missing values or values outside the scale range
```

```
## (`geom_point()`).
```




```
p + facet_wrap(~Country) +
xlab("GDP per capita") +
ylab("Electricity consumption per capita")
```

```
## Warning: Removed 1181 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

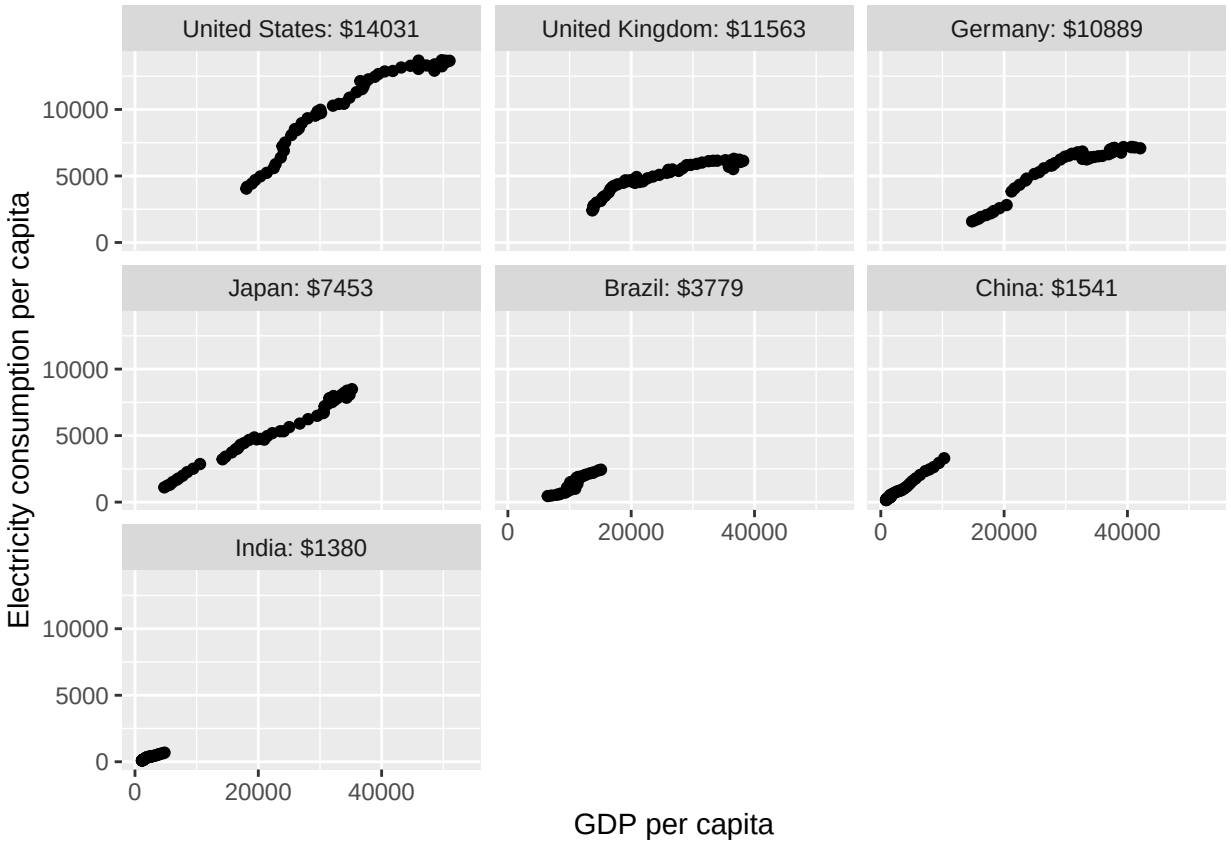


```
ordered_countries <- df %>%
  group_by(Country) %>%
  summarize(mean = round(mean(gdp_per_capita))) %>%
  arrange(desc(mean)) %>%
  mutate(labels = str_c(Country, ": $", mean))
country.labs <- ordered_countries$labels
names(country.labs) <- ordered_countries$Country

df_ordered <- df %>%
  mutate(Country = factor(Country, levels=ordered_countries$Country))

ggplot(df_ordered, aes(x=gdp_per_capita, y=Electricity_consumption_per_capita)) +
  geom_point() +
  facet_wrap(~Country,
             labeller=labeller(Country = country.labs)) +
xlab("GDP per capita") +
ylab("Electricity consumption per capita")
```

```
## Warning: Removed 1181 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



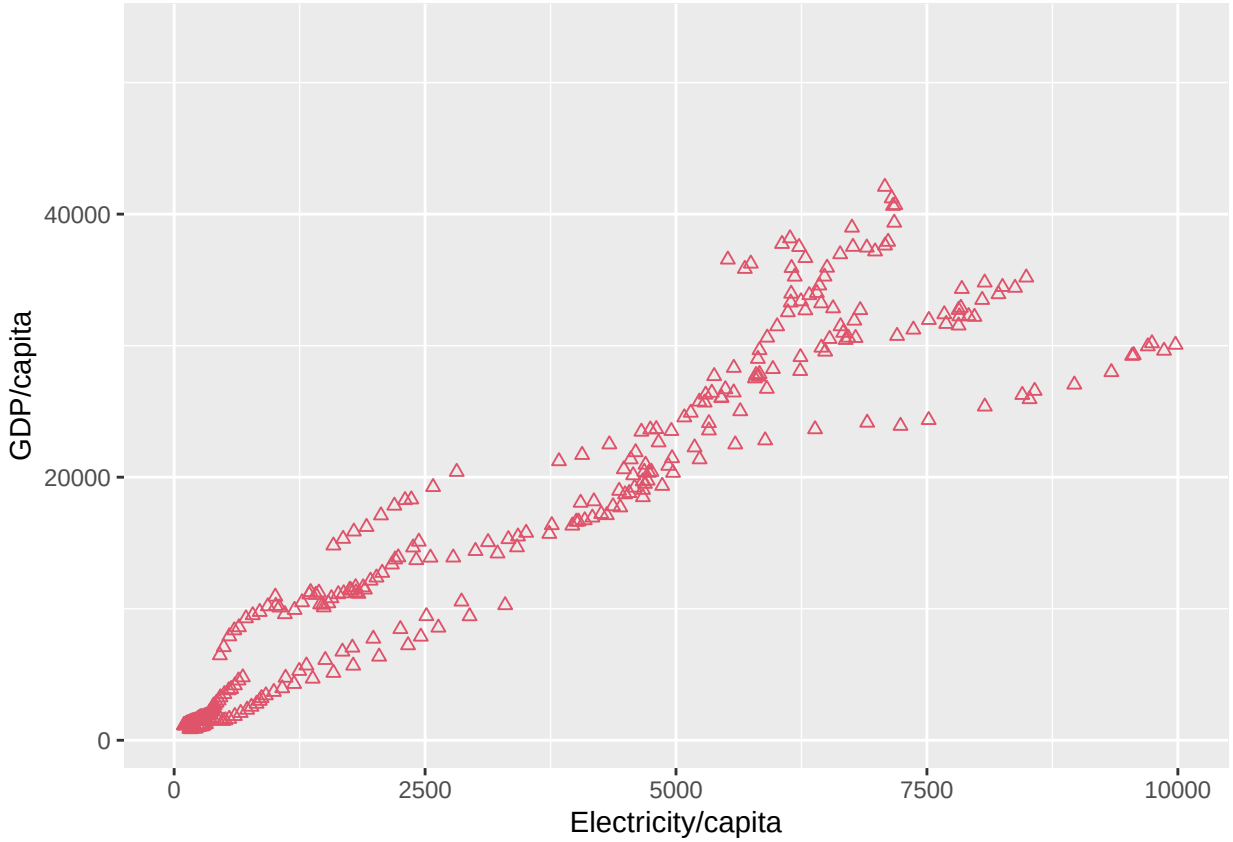
Shapes and colors

```
dfs <- subset(df, Country %in% c("Germany", "India", "China", "United States"))
var1 <- "Electricity_consumption_per_capita"
var2 <- "gdp_per_capita"
```

```
name1 <- "Electricity/capita"
name2 <- "GDP/capita"
ggplot(df, aes_string(x=var1, y=var2)) +
geom_point(color=2, shape=2) +
xlim(0, 10000) +
xlab(name1) +
ylab(name2)
```

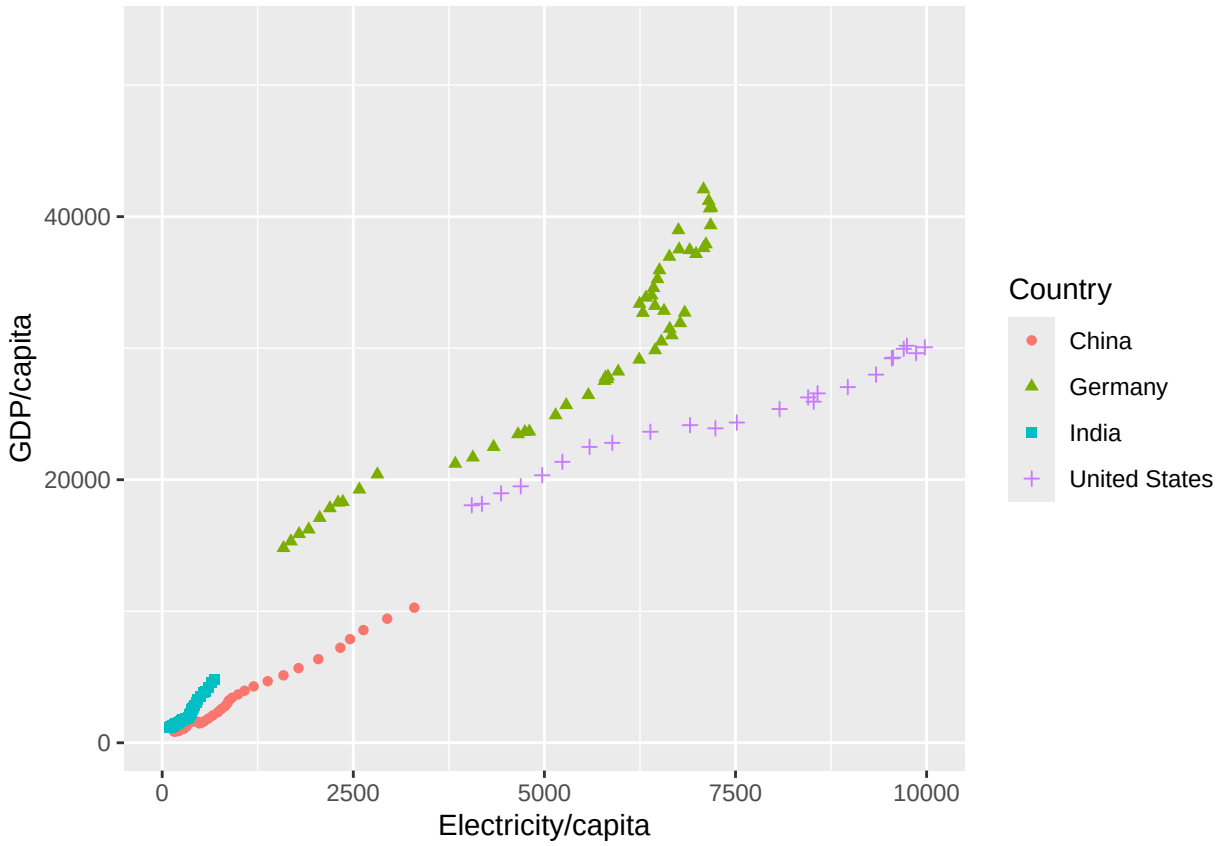
```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: Removed 1209 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
ggplot(dfs, aes_string(x=var1, y=var2)) +
geom_point(aes(color=Country, shape=Country)) +
xlim(0, 10000) +
xlab(name1) +
ylab(name2)
```

```
## Warning: Removed 706 rows containing missing values or values outside the scale range
## (`geom_point()`).
```


GDP/capita
40000
20000
0
0
2500
5000
7500
10000
Electricity/capita
Country
China
Germany
India
United States