Data Visualization with R
Ggplot2 tutorial (part 1)

Ariane Ducellier

University of Washington - Fall 2023

What is ggplot2?

Ggplot2 is the graphics package from the tidyverse, a collection of R packages designed for data science.

There is a base graphics package in R, which is present in the default version of R.

However, ggplot2 gives users a lot more flexibility and control over their visualizations.

Main concepts of ggplot2

Ggplot2 is based on layers:

- A first layer to describe the dataset.

- Layers describing the objects representing the data (dots, lines, bars, etc.).

- Additional objects describing the graphic itself (coordinates, scales, fonts, etc.).

Example: Histograms
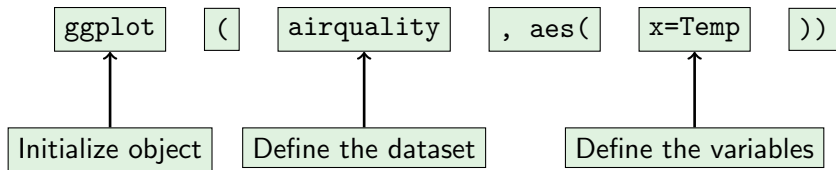
Built-in R graphics package:

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

Quick plot using ggplot2:

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

Ggplot2 command structure

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

This command does not plot anything.

| ggplot | (| airquality | , aes(| x=Temp |)) |

Initialize object   Define the dataset   Define the variables

We need to add a command to explain the kind of object that we want to plot:

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

We can use bar plots to visualize one categorical variable:

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

The height of the bar is proportional to the number of cases in each group.

Bar plots

Or a combination of a categorical variable and a continuous variable:

```
ggplot(RetailSales, aes(x=Month, y=Sales)) +
geom_bar(stat="identity")
```

Using stat = "identity" tells ggplot2 to sum the values for each group (Month) and plot bars proportional to the sums.

For each layer that we want to add on our plot, we add the corresponding object:

```
ggplot(df_hum, aes(x=month, y=Vancouver)) +
geom_boxplot()
```

Scatter plots and line plots

The relationship between two continuous variables can be visualize with a scatter plot or a line plot:

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

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

Changing histogram defaults

Modify the number of bins:

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

Modify the filling and the color:

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

Adding aesthetics to the plot

Add title and axis labels to the histogram:

```
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))
```

Adding aesthetics to the boxplot

Add labels to the box plot:

```
ggplot(df_hum, aes(x=month, y=Vancouver)) +
geom_boxplot(color=1, fill=3) +
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=2),
axis.title.y=element_text(size=15, color=2))
```