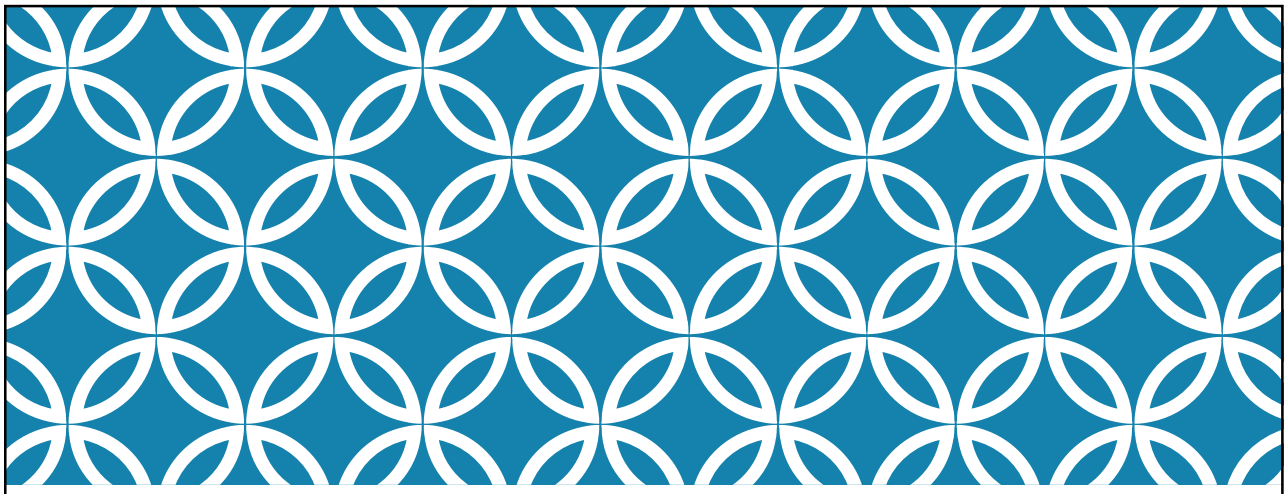


Data Visualization

Session 3

Stephan Kadauke

1

1

Session	Instructor
Instructor Introductions, Introduction to technology	Amrom Obstfeld
Introduction to R and RStudio	Joe Rudolf
Reproducible Reporting	Joe Rudolf
Data Visualization	Stephan Kadauke
Data Transformation	Amrom Obstfeld
Statistical Analysis	Dan Herman
Advanced Reporting	Patrick Mathias

2

2

Goals

1. Appreciate the importance of visualization for understanding data
2. Learn how to use `ggplot2` to visualize data

Objectives

1. Create a basic visualization using a simple **template**
2. Define "**aesthetic mapping**" and explain how aesthetic mappings relate variables of a data frame to features of graphic markings on a plot
3. Write the code to specify a type of plot and fine tune its appearance using "**geom**" functions
4. Explain how to add **layers** to a `ggplot` object to create complex and highly customized visualizations

3

3

`covid_testing`

covid_testing

Filter

	mrn	first_name	last_name	gender	pan_day	test_id	clinic_name	result
1	5001412	jhezane	westerling	female	4	covid	inpatient ward a	negative
2	5000533	penny	targaryen	female	7	covid	clinical lab	negative
3	5009134	grunt	rivers	male	7	covid	clinical lab	negative
4	5008518	melisandre	swyft	female	8	covid	clinical lab	negative
5	5008967	rolley	karstark	male	8	covid	emergency dept	negative
6	5011048	megga	karstark	female	8	covid	oncology day hosp	negative
7	5000663	ithoke	targaryen	male	9	covid	clinical lab	negative
8	5002158	ravella	frey	female	9	covid	emergency dept	negative
9	5003794	styr	tyrell	male	9	covid	clinical lab	negative
10	5004706	wynafryd	seaworth	male	9	covid	clinical lab	negative
11	5008115	patrek	frey	male	9	covid	clinical lab	negative
12	5009309	maege	sand	female	9	covid	medical center	negative
13	5008943	myria	rivers	female	9	covid	picu	negative

Showing 1 to 14 of 15,524 entries, 17 total columns

4

4

Your Turn 1

Consider the `covid_testing` data frame.

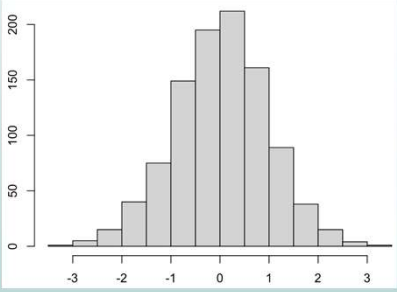
What do you think plot would look like in which:

- the x-axis represents `pan_day` (day of the pandemic), and
- the y-axis represents the number of tests that were performed on that day?

5

5

Your Turn 2



What is the name of this kind of plot?

Type the answer into the chat!

6

6

Your Turn 3

Type the following code in the RStudio console to make a graph.

Pay attention to the spelling, capitalization, and parentheses!

```
ggplot(data = covid_testing) +
  geom_histogram(mapping = aes(x = pan_day))
```

7

7

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

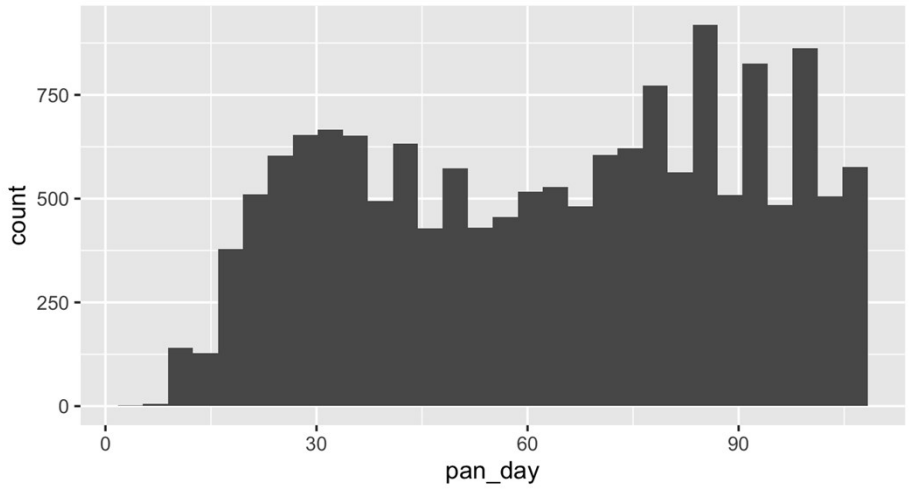
Often times, things that look like an error in R are actually just a message.

R lets you know that when you ask it to draw a histogram you should tell it how wide each bin should be, because this affects the granularity of the data displayed.

```
ggplot(data = covid_testing) +
  geom_histogram(mapping = aes(x = pan_day))
```

8

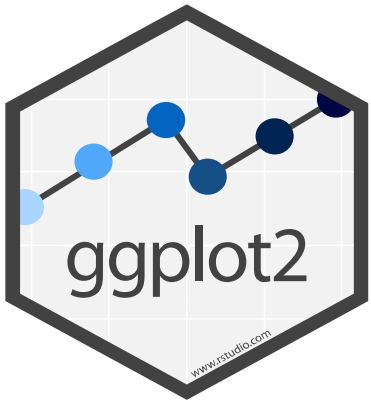
8




```
ggplot(data = covid_testing) +
  geom_histogram(mapping = aes(x = pan_day))
```

9

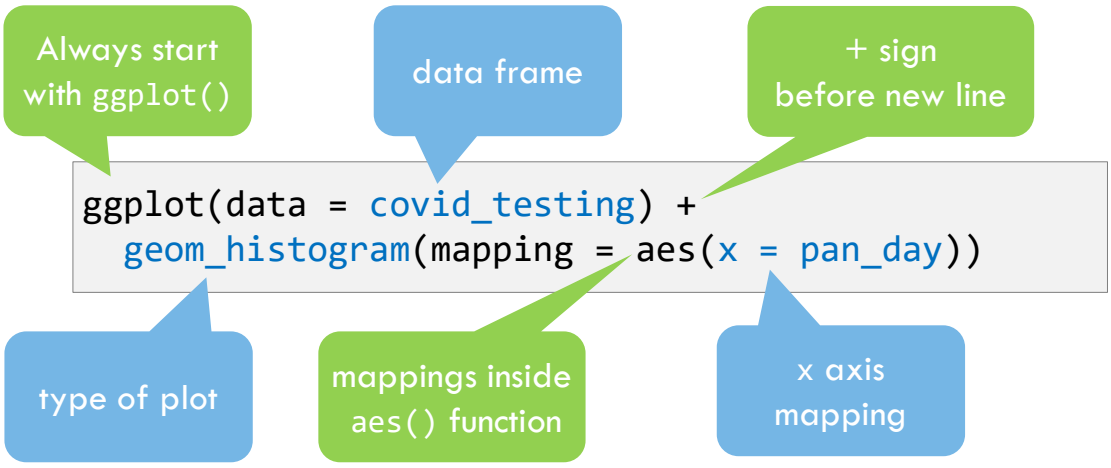
9



10

10

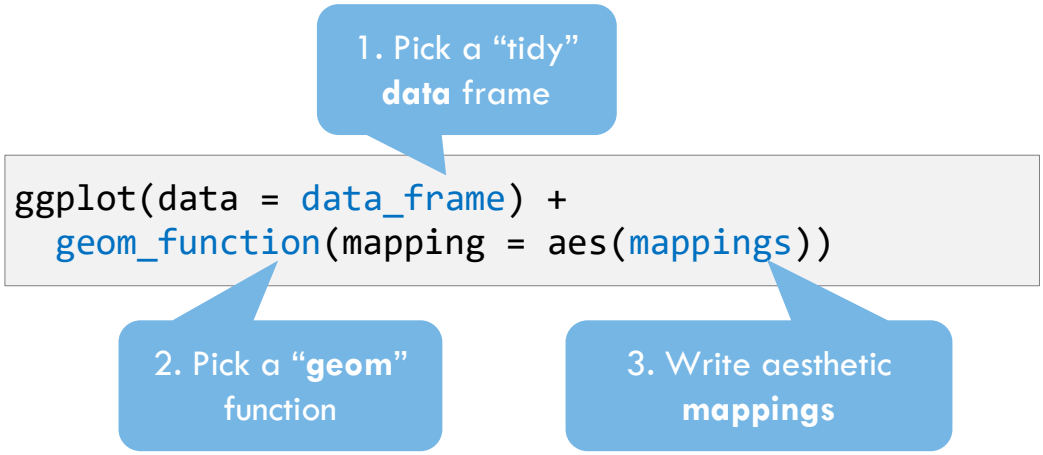
ggplot()



11

11

To make **any** kind of graph:



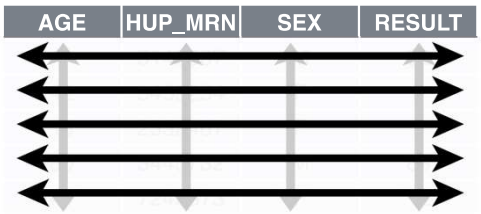
12

1. Pick a "tidy" **data** frame

```
ggplot(data = data_frame) +
  geom_function(mapping = aes(mappings))
```

13

1. Pick a "Tidy" Data Frame



A data set is **tidy** if:

1. Each **variable** is in its own **column**
2. Each **observation** is in its own **row**
3. Each **value** is in its own **cell**

Wickham H. **Tidy Data.** *J Stat Soft* 2014.

14

1. Pick a "tidy" **data** frame

```
ggplot(data = data_frame) +
  geom_function(mapping = aes(mappings))
```

2. Pick a "**geom**" function

15

2. Pick a "Geom" Function

`geom_histogram()`

`geom_dotplot()`

`geom_bar()`

`geom_boxplot()`

`geom_point()`

`geom_line()`

16

16

1. Pick a "tidy" **data** frame

```
ggplot(data = data_frame) +
  geom_function(mapping = aes(mappings))
```

2. Pick a "**geom**" function

3. Write aesthetic **mappings**

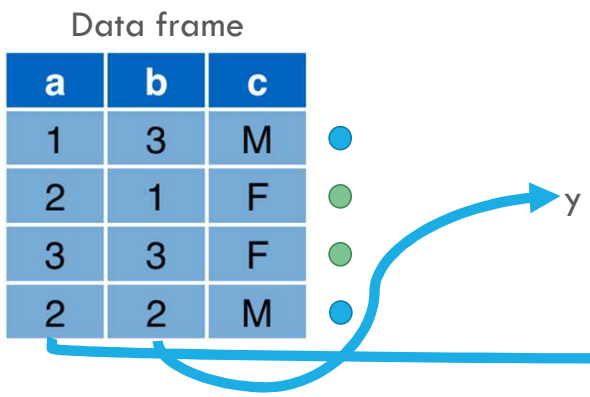
17

3. Write Aesthetic Mappings

```
aes(x = a, y = b, color = c)
```

Data frame

a	b	c
1	3	M
2	1	F
3	3	F
2	2	M

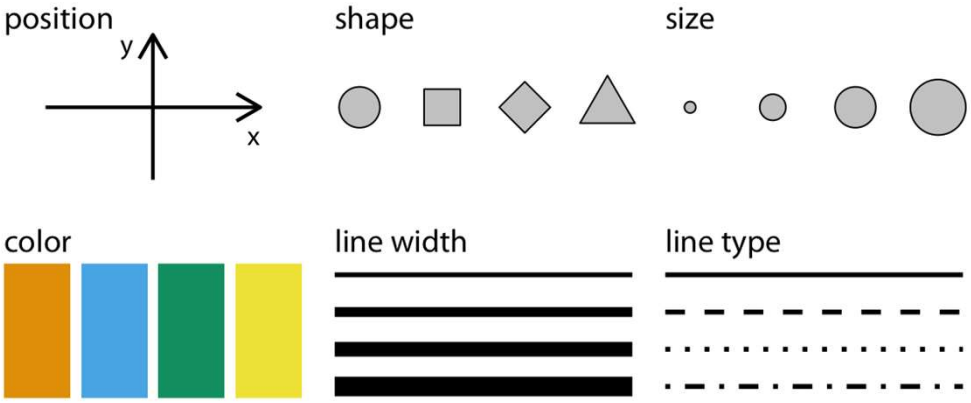


Graph

18

18

Aesthetics



To make **any** kind of graph:

1. Choose a "tidy" **data** frame

```
ggplot(data = data_frame) +
  geom_function(mapping = aes(mappings))
```

2. Pick a "**geom**" function

3. Write aesthetic **mappings**

Your Turn 4

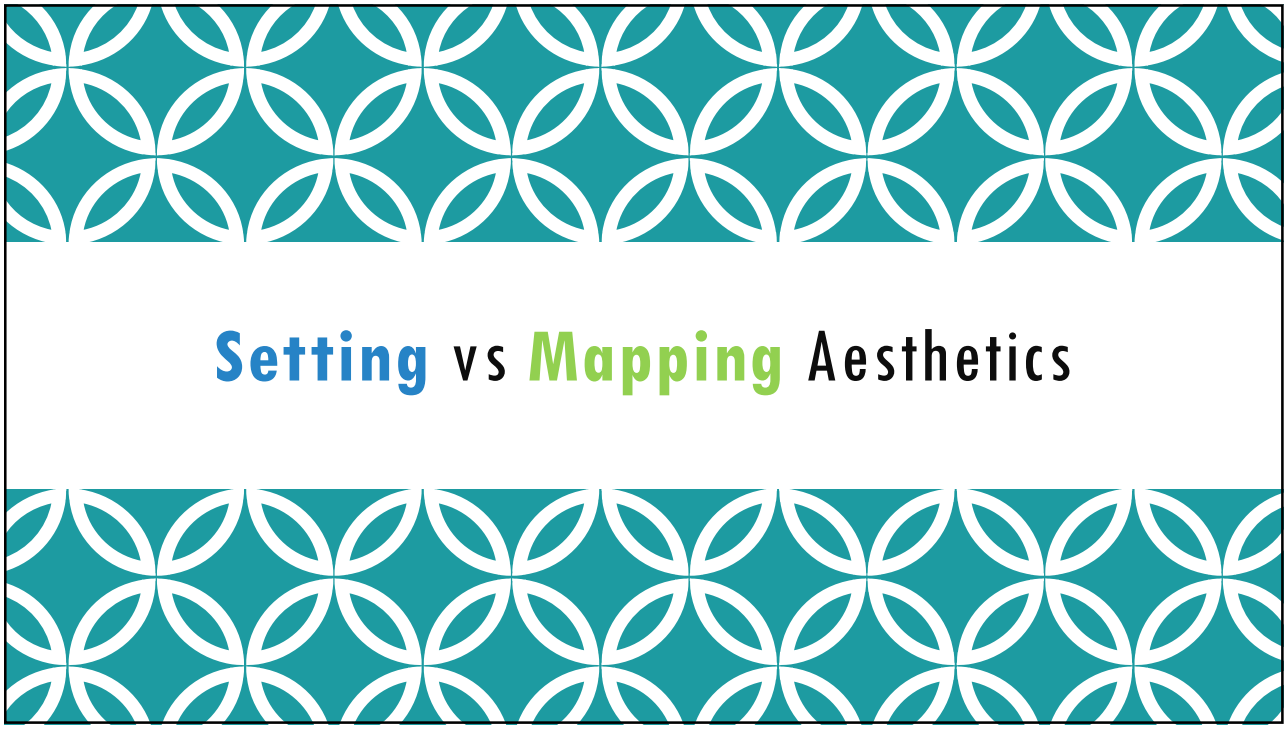
Open `03 - Visualize.Rmd`. Work through the exercises of the section titled "Your Turn 4".

Stop when it says "Stop Here".

Click "yes" when you're done!

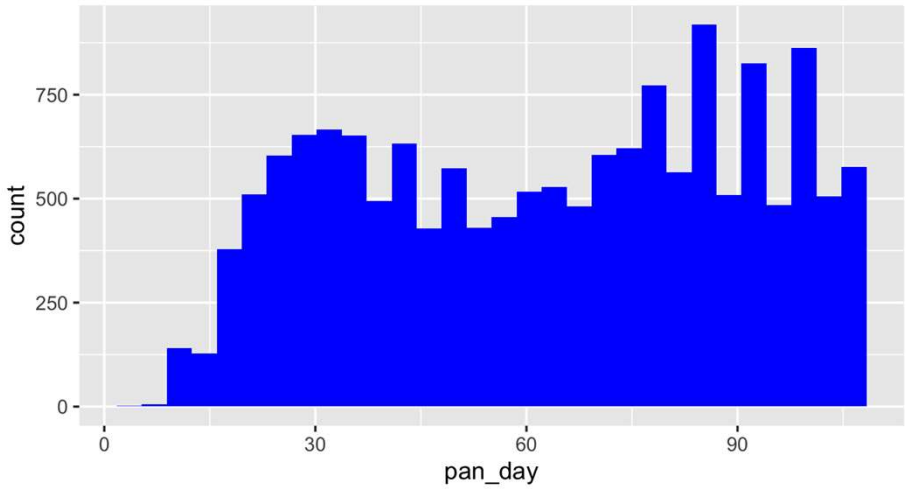
21

21



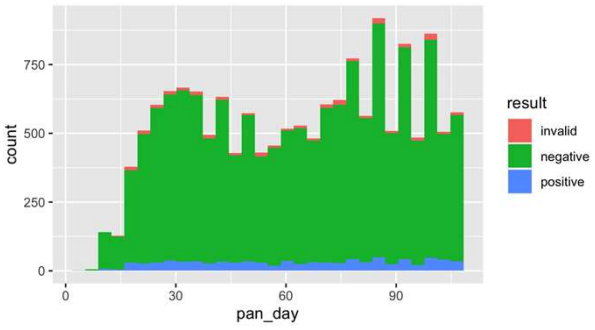
22

How would you make this plot?



23

23

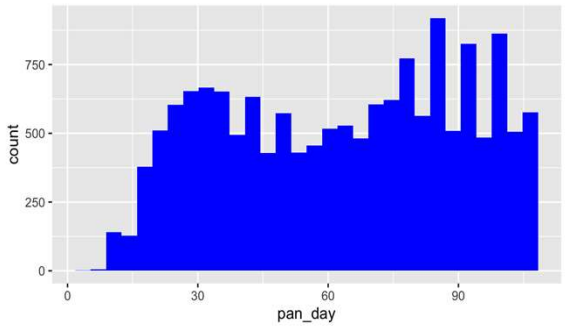


Inside of `aes()`:
map an aesthetic
to a **variable**

```
ggplot(data = covid_testing) +
  geom_histogram(mapping = aes(x = pan_day, fill = result))
```

24

24



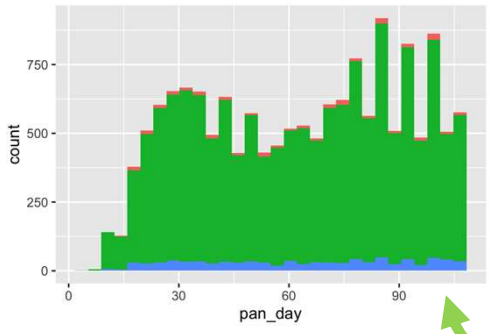
Outside of `aes()`:
set an aesthetic to a **value**

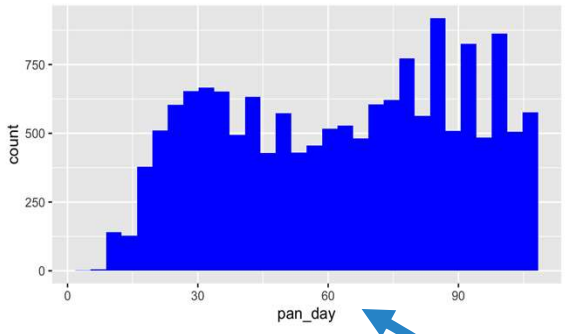
color name in "quotes"

```
ggplot(data = covid_testing) +
  geom_histogram(mapping = aes(x = pan_day), fill = "blue")
```

25

25






```
ggplot(data = covid_testing) +
  geom_histogram(mapping = aes(x = pan_day, fill = result))
```

```
ggplot(data = covid_testing) +
  geom_histogram(mapping = aes(x = pan_day), fill = "blue")
```

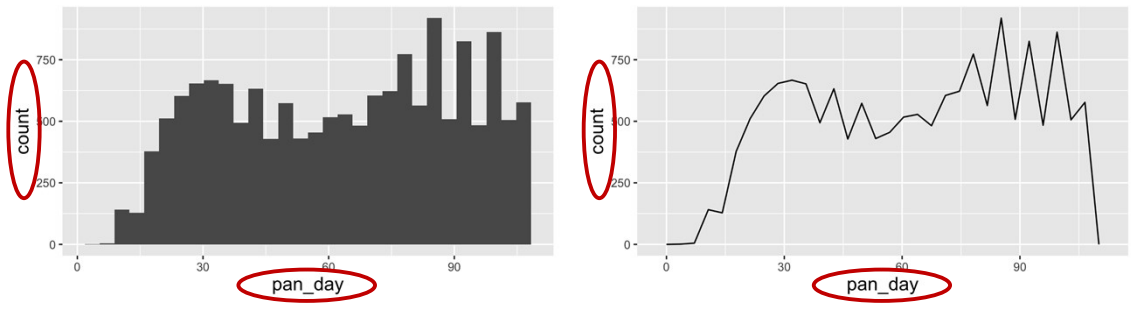
26

26

Geom Functions

27

How are these plots similar?

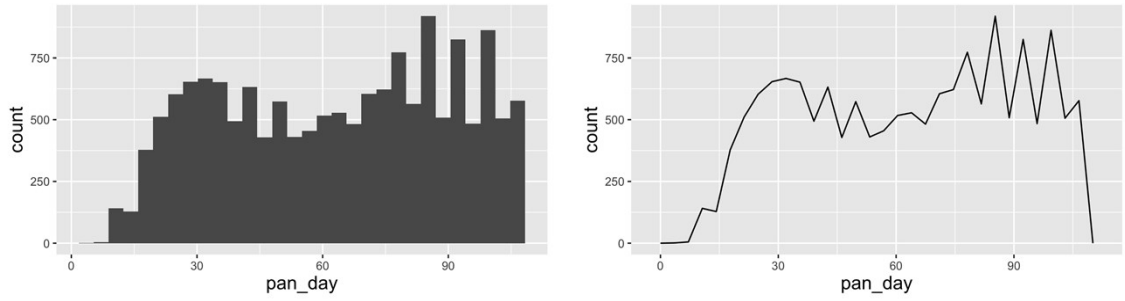


Same: x axis, y axis, data

28

28

How are these plots different?



Different **geometric object** ("geom") used to represent the data

29

29

Your Turn 5

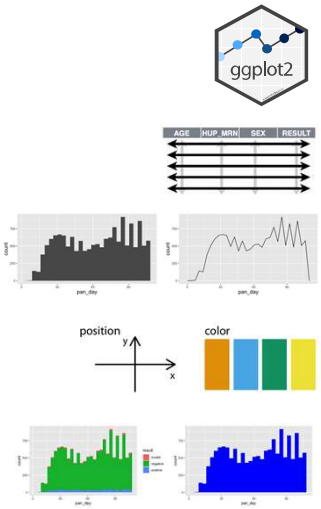
Return to `03-Visualize.Rmd`. Work through the exercises of the section titled "Your Turn 5."

Click "yes" when you're done!

30

30

Recap



ggplot2 is a package that provides a **grammar of graphics**. You can create **any type of plot** using a simple template to which you provide:

1. A **tidy data frame**, in which each **variable** is in its own **column**, each **observation** is in its own **row**, each **value** is in its own **cell**;

2. A **geom function**, which tells R what kind of plot to make; and

3. **Aesthetic mappings**, which tell R how to represent data as graphical markings on the plot.

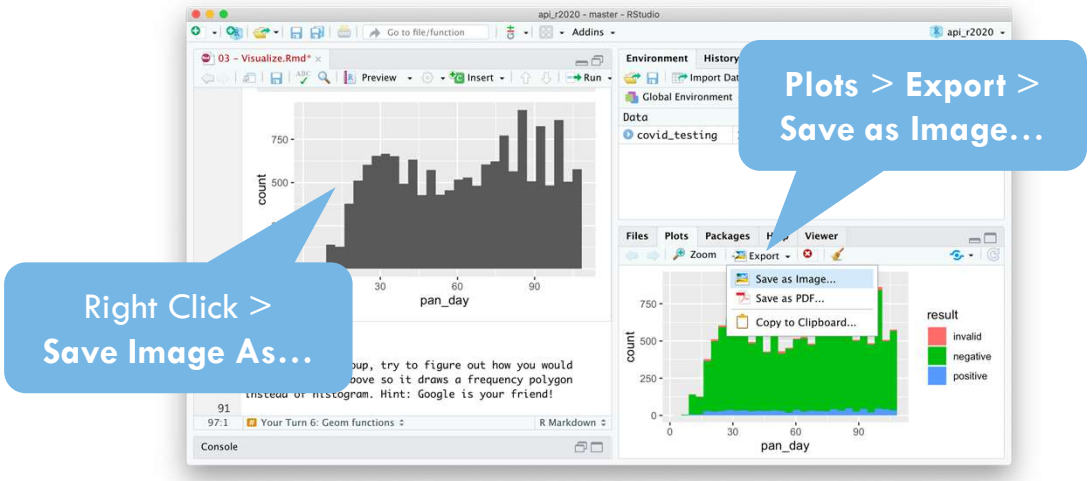
Aesthetics can be **mapped** to a variable or **set** to a constant value.

31

What Else?

32

Manually saving plots

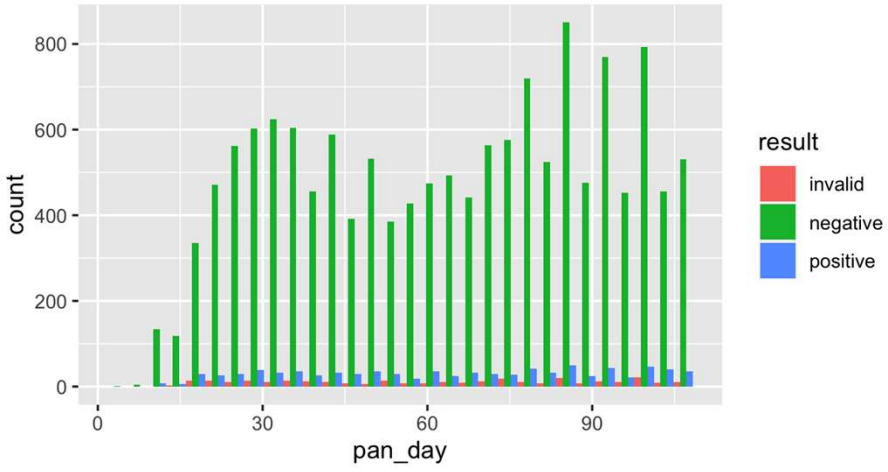


33

33

Position adjustments

How overlapping objects are arranged

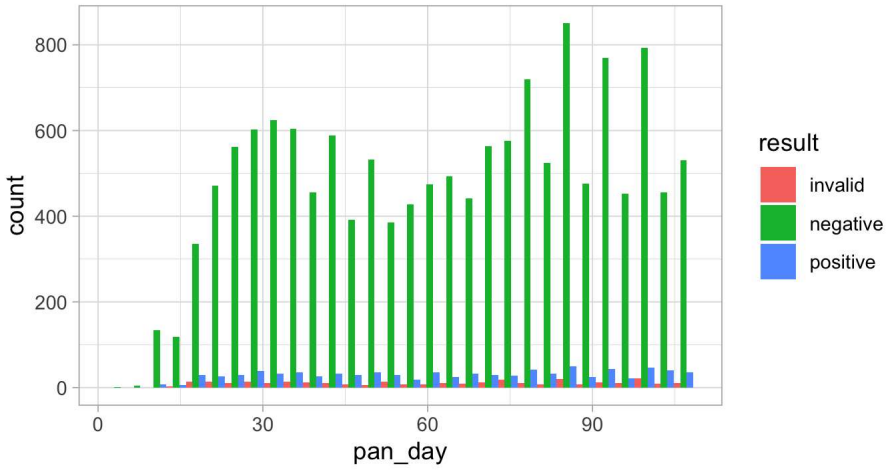


34

34

Themes

Visual appearance of non-data elements

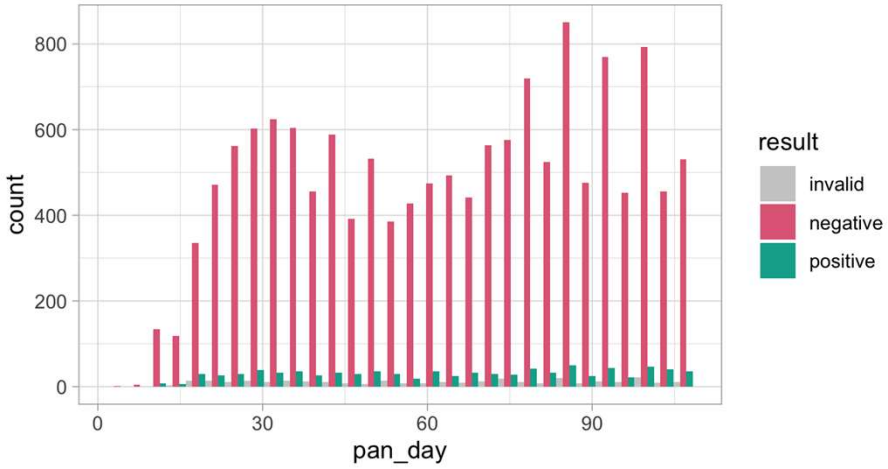


35

35

Scales

Customize color scales and other mappings

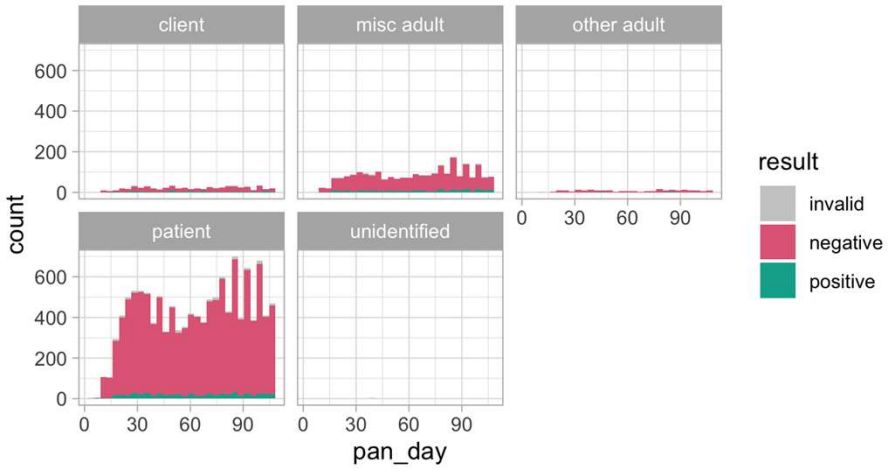


36

36

Facets

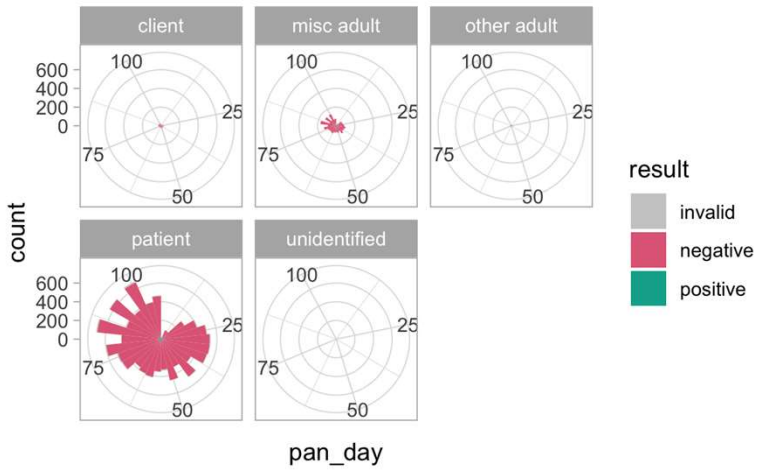
Subplots that display subsets of the data



37

37

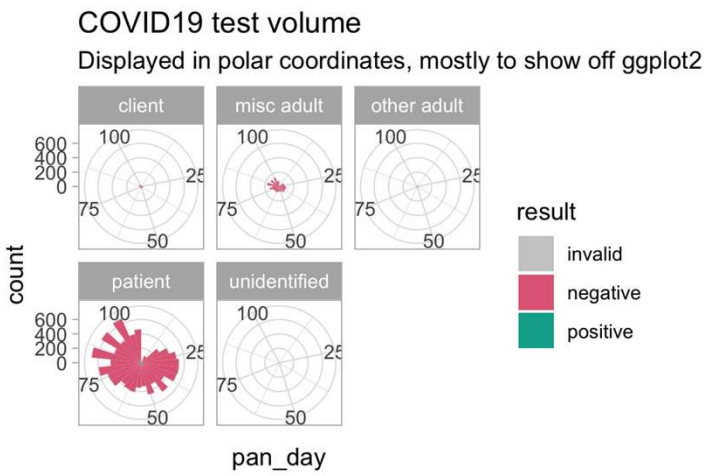
Coordinate systems



38

38

Titles and captions



39

39

```
ggplot(data = data_frame) +
  geom_function(mapping = aes(mappings)) +
  theme_function +
  scale_function +
  facet_function +
  coordinate_function +
  ...
```

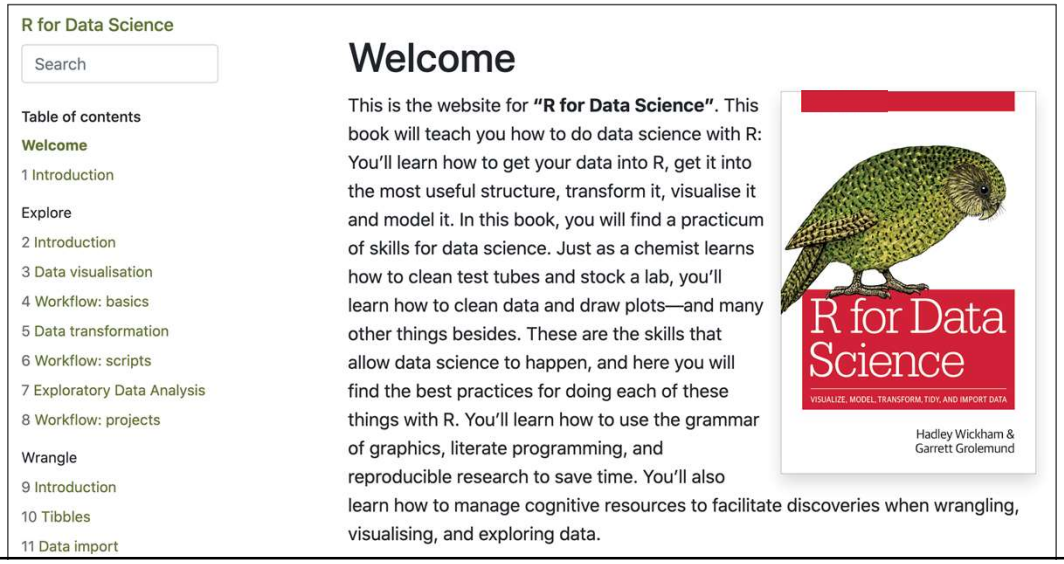
Required

Optional

40

40

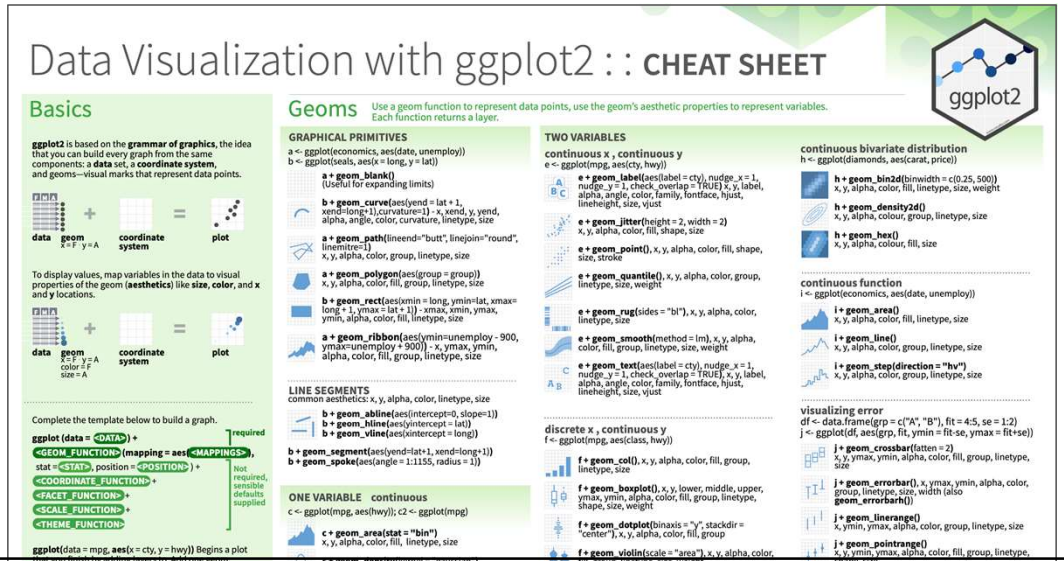
https://r4ds.had.co.nz/



41

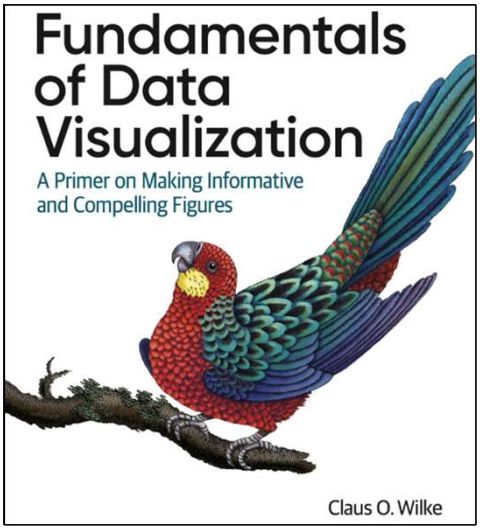
41

https://www.rstudio.com/resources/cheatsheets/



42

42



https://clauswilke.com/dataviz/

43

43

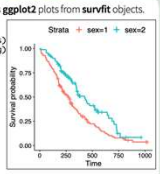
Creating Survival Plots

Informative and Elegant with *survminer*

Survival Curves

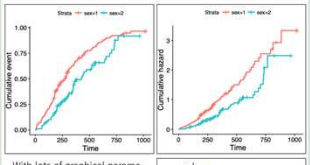
The **ggsurvplot()** function creates **ggplot2** plots from **survfit** objects.

```
library("survival")
fit <- survfit(Surv(time,status)
               ~ sex, data = lung)

class(fit)
## [1] "survfit"

library("survminer")
ggsurvplot(fit, data = lung)
```



Use the **fun** argument to set the transformation of the survival curve. E.g. **"event"** for cumulative events, **"cumhaz"** for the cumulative hazard function or **"pct"** for survival probability in percentage.

```
ggsurvplot(fit, data = lung, fun = "event")
ggsurvplot(fit, data = lung, fun = "cumhaz")
```

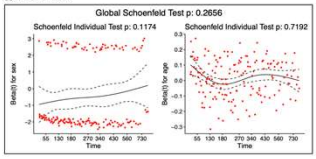


With lots of graphical parameters you have full control over look and feel of the survival

Diagnostics of Cox Model

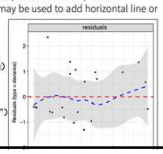
The function **cox.zph()** from **survival** package may be used to test the proportional hazards assumption for a Cox regression model fit. The graphical verification of this assumption may be performed with the function **ggcoxzph()** from the **survminer** package. For each covariate it produces plots with scaled Schoenfeld residuals against the time.

```
library("survival")
fit <- coxph(Surv(time, status) ~ sex + age, data = lung)
ftest <- cox.zph(fit)
ftest
##          rho chisq     p
## sex    0.1236 2.452 0.117
## age   -0.0275 0.129 0.719
## GLOBAL     NA 2.651 0.266
library("survminer")
ggcoxzph(ftest)
```



The function **ggcoxdiagnostics()** plots different types of residuals as a function of time, linear predictor or observation id. The type of residual is selected with **type** argument. Possible values are "martingale", "deviance", "score", "schoenfeld", "dfbeta", "dfbetas", and "scaledsch".
The **ox.scale** argument defines what shall be plotted on the OX axis. Possible values are "linear.predictions", "observation.id", "time".
Logical arguments **hline** and **sline** may be used to add horizontal line or smooth line to the plot.

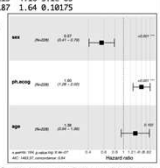
```
library("survival")
library("survminer")
fit <- coxph(Surv(time, status) ~
             sex + age, data = lung)

ggcoxdiagnostics(fit,
     type = "deviance",
     ox.scale = "linear.predictions")

ggcoxdiagnostics(fit,
```



Summary of Cox Model

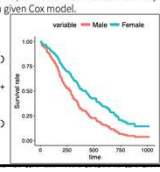
The function **ggforest()** from the **survminer** package creates a forest plot for a Cox regression model fit. Hazard ratio estimates along with confidence intervals and p-values are plotter for each variable.

```
library("survival")
library("survminer")
lung$age <- ifelse(lung$age > 70, ">70","<= 70")
fit <- coxph( Surv(time, status) ~ sex + ph.ecog + age, data = lung)
fit

## Call:
## coxph(formula = Surv(time, status) ~ sex+ph.ecog+age, data=lung)
##
##             coef exp(coef) se(coef)     z        p
## sex      -0.567     0.567    0.168 -3.37 0.00075
## ph.ecog  0.470     1.600    0.113  4.16 3.1e-05
## age>70   0.307     1.359    0.187  1.64 0.10175
##
## Likelihood ratio test=31.6  on
## n= 227, number of events= 164

ggforest(fit)
```



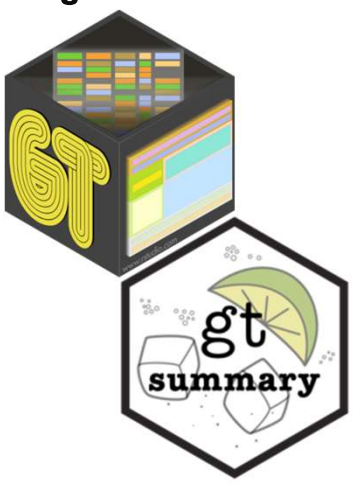
The function **ggadjustedcurves()** from the **survminer** package plots Adjusted Survival Curves for Cox Proportional Hazards Model. Adjusted Survival Curves show how a selected factor influences survival estimated from a Cox model.
Note that these curves differ from Kaplan Meier estimates since they present expected ssurvival based on given Cox model.

```
library("survival")
library("survminer")

lung$sex <- ifelse(lung$sex == 1,
                   "Male", "Female")

fit <- coxph(Surv(time, status) ~
             sex + ph.ecog + age +
             strata(sex),
             data = lung)

ggcoxadjustedcurves(fit, data=lung)
```



44

A grammar for **tables**



Variable	N	Drug A, N = 98[1]	Drug B, N = 102[1]	p-value[2]
Age	189	46 (37, 59)	48 (39, 56)	0.7
Grade	200			0.9
I		35 (36%)	33 (32%)	
II		32 (33%)	36 (35%)	
III		31 (32%)	33 (32%)	
Tumor Response	193	28 (29%)	33 (34%)	0.6

[1] Statistics presented: median (IQR); n (%)

[2] Statistical tests performed: Wilcoxon rank-sum test; chi-square test of independence

45

Goals

1. Appreciate the importance of visualization for understanding data
2. Learn how to use ggplot2 to visualize data

Objectives

1. Create a basic visualization using a simple **template**
2. Define "**aesthetic mapping**" and explain how aesthetic mappings relate variables of a data frame to features of graphic markings on a plot
3. Write the code to specify a type of plot and fine tune its appearance using "**geom**" functions
4. Explain how to add **layers** to a ggplot object to create complex and highly customized visualizations

46

46