Ling 105
Sounds of Language
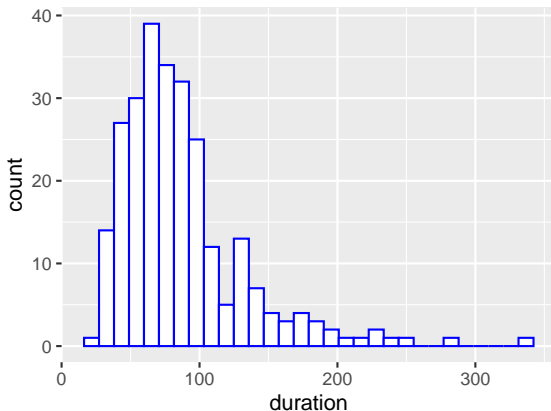
Thursday, October 24, 2024

Kevin Ryan

Today

- Visualization (**ggplot2**, part of **tidyverse**)
- Reading: *R for Data Science (2e)* chapter 1
- Data for today: **vowels2.txt**
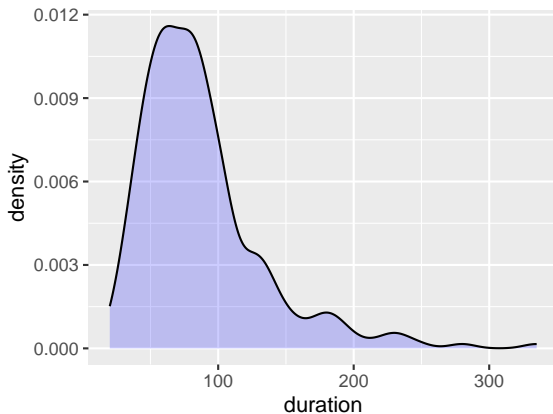- R code from class is posted on Canvas

Histogram

- Duration of [i] for speaker 1 ($\bar{x} = 87$ ms)
- ggplot(aes(x = duration)) +
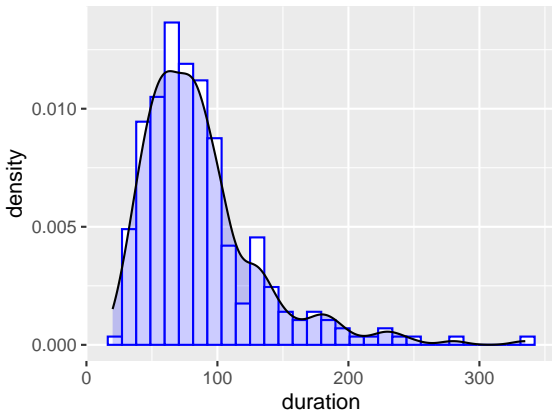 geom_histogram(color = "blue", fill = "white")

Density plot

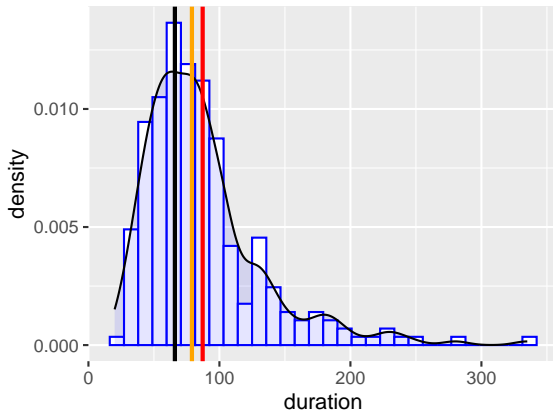- geom_density(fill = "blue", alpha = 0.2)

Layers

- Add both geoms with + (order matters)
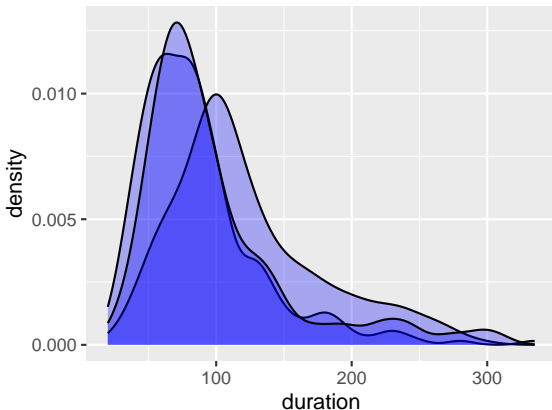- But the histogram must be rescaled: y = after_stat(density)

A third layer

- Add lines for the **mean**, **median**, and **mode**
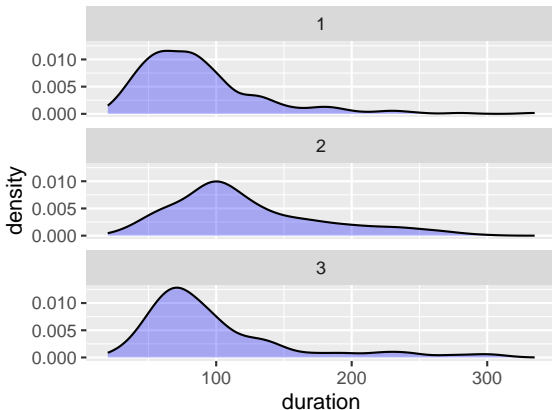- Which is which?
- geom_vline(xintercept = ...)

Separating speakers (method 1)

- Redefine the data frame to include the first three speakers
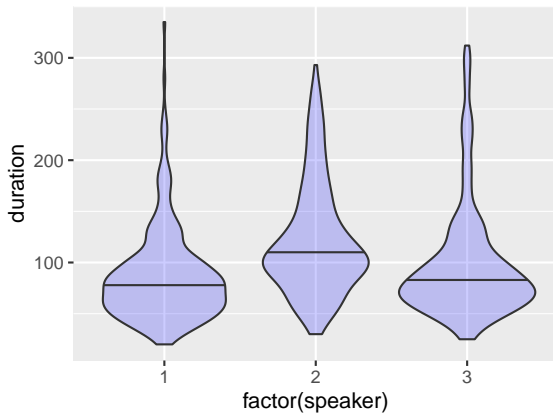- Add to aes: group = factor(speaker)

Separating speakers (method 2)

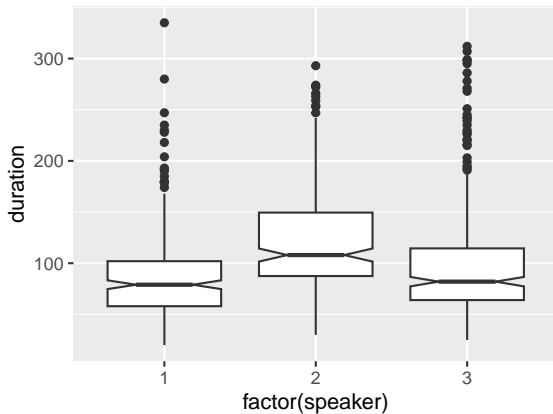- facet_wrap($\sim$ factor(speaker), ncol = 1)
- (Not part of aes)
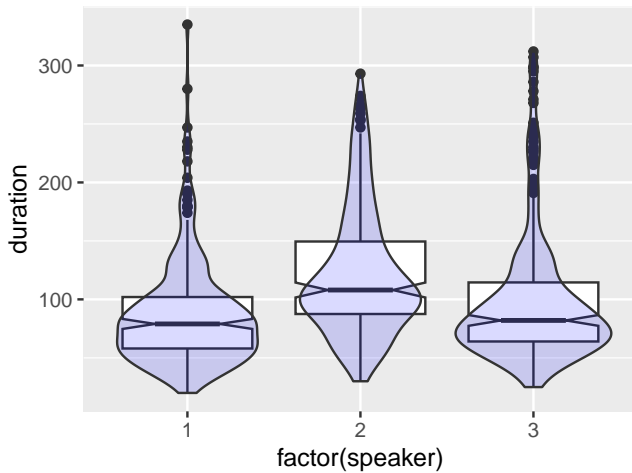
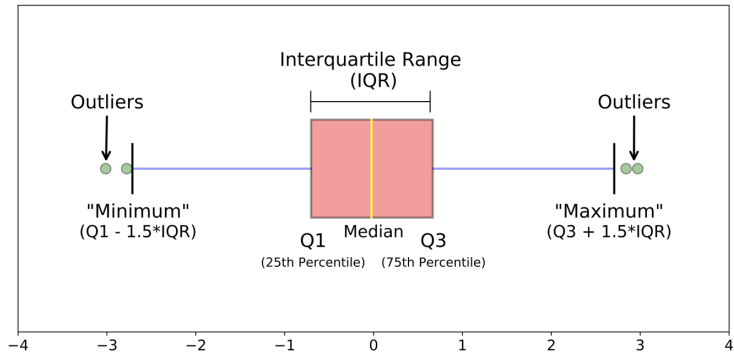# Violin/bean plot

- geom_violin(...)

Box plot

- geom_boxplot(), here with notch = TRUE, shows less detail, but summarizes key properties better
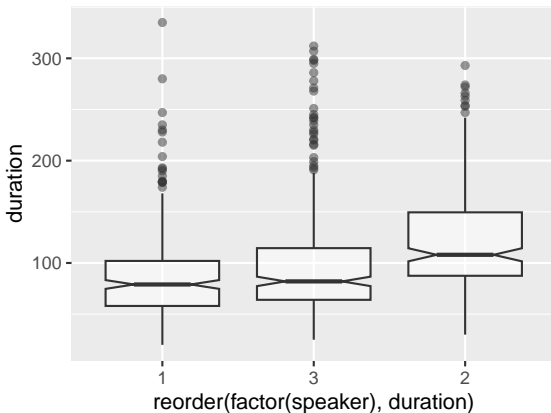
Box + bean, compared

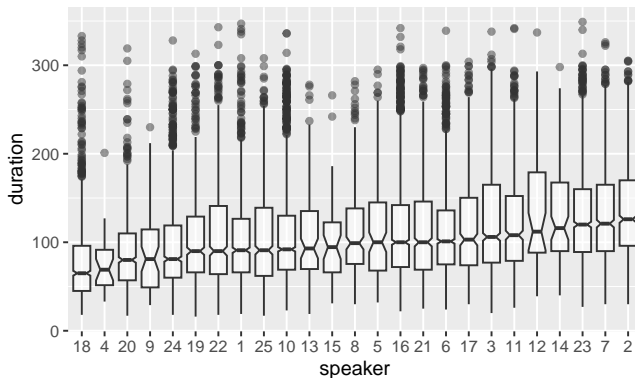# Reading a box plot

(by Michael Galarnyk)

Reordered box plot

- Reorder bars by median for easier comparisons
- x = reorder(factor(speaker), duration, FUN = median)
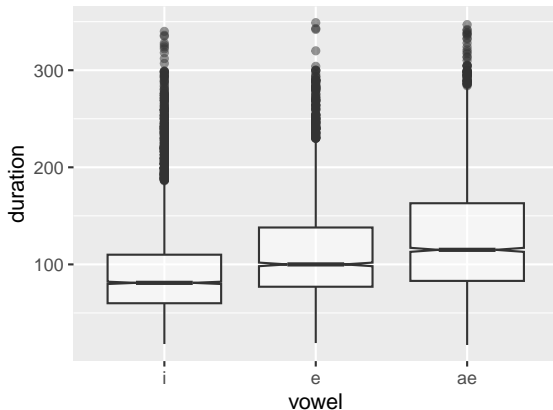
Reordered box plot

- 25 speakers
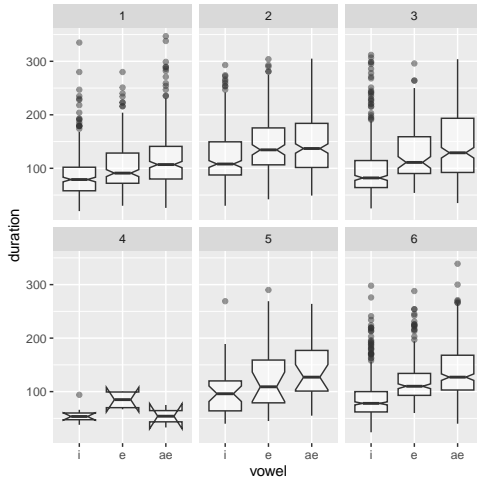- Fix the label: labs(x = "speaker")

Lower vowels are longer

- All else (e.g. tenseness) equal, progressively lower vowels tend to be progressively longer
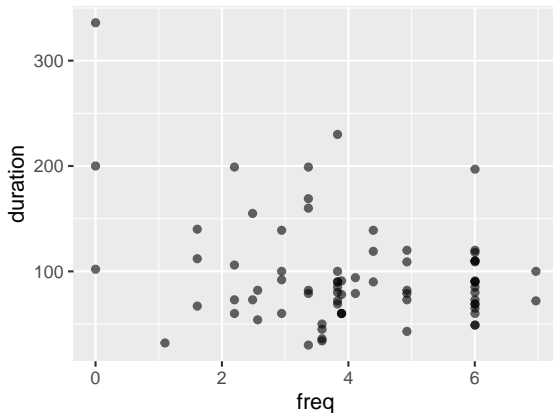- Why?

Lower vowels are longer

- Consistent across speakers (what's up with 4?)

Scatterplot

- As frequency increases, duration tends to decrease
- Here, an arbitrary small sample of [i]
- geom_point(...)

Adding a smoother

- Adding geom_smooth()
- Uses a method called LOESS (curvy fit)