

Lecture 6: Multivariate visualizations

Stats 32: Introduction to R for Undergraduates

Harrison Li

April 18, 2024

Agenda

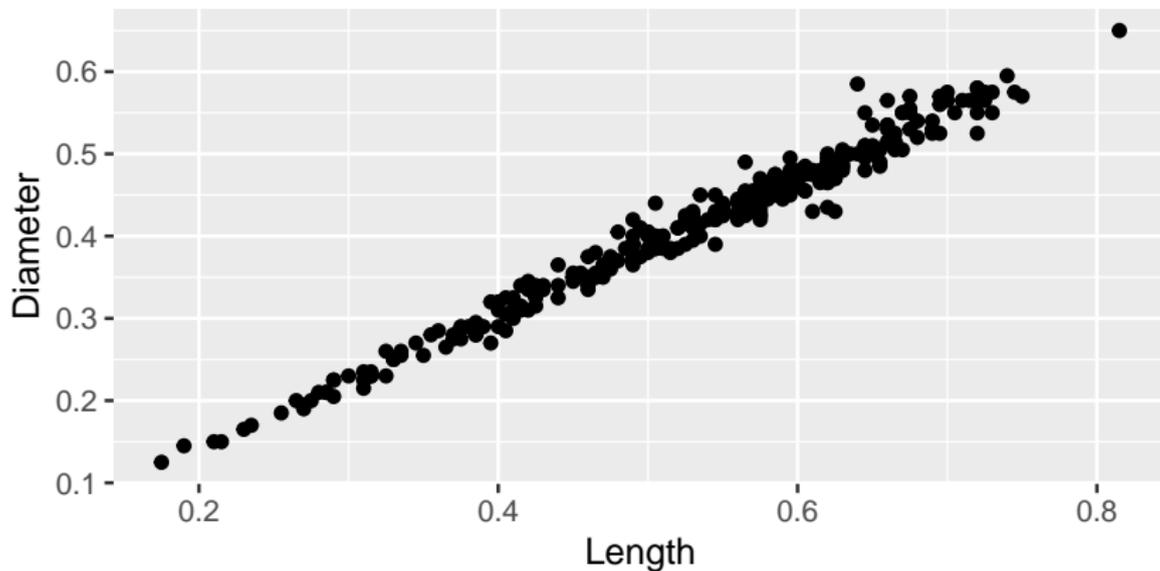
- 1 Scatterplots
- 2 Line plots
- 3 Two categorical variables
- 4 Summary of main plot types
- 5 Layers
- 6 Faceting

Reading: Sections 2.2-2.4, 2.6, 2.9

Scatterplots

Scatterplots

A **scatterplot** consists of points on a set of two perpendicular axes. It is useful for visualizing relationships between two *quantitative* variables.



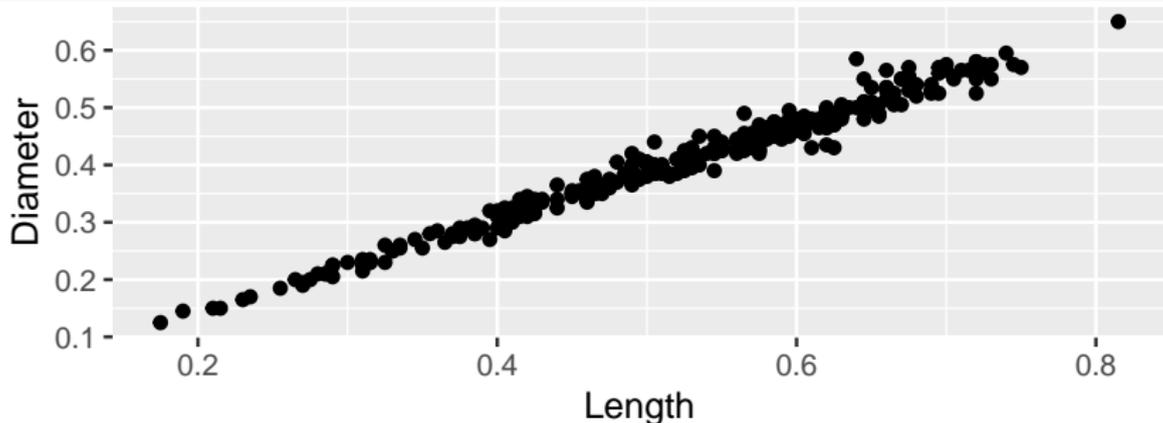
geom_point()

`geom_point()` is the standard geom for scatterplots.

Let's look at abalone Diameter vs. Length:

```
library(tidyverse)
abalone <- read_csv("abalone.csv")

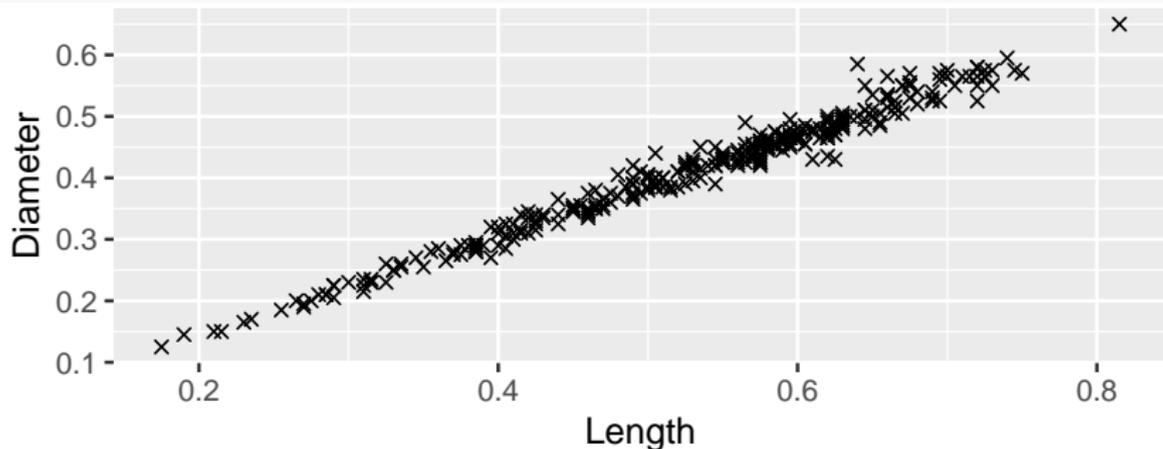
abalone %>%
  ggplot(aes(x=Length, y=Diameter)) +
  geom_point()
```



geom_point()

We can change the marker type to X's (this is a static aesthetic change, so goes outside `aes()` in the `shape` argument):

```
abalone %>%  
  ggplot(aes(x=Length, y=Diameter)) +  
  geom_point(shape=4)
```



A full list of shapes can be found in the help page for `geom_point()`.

aes() revisited

Now let's use `aes()` to set the point color *based on a data variable*.

For instance, we might want to color points based on whether their height is above or below average.

aes() revisited

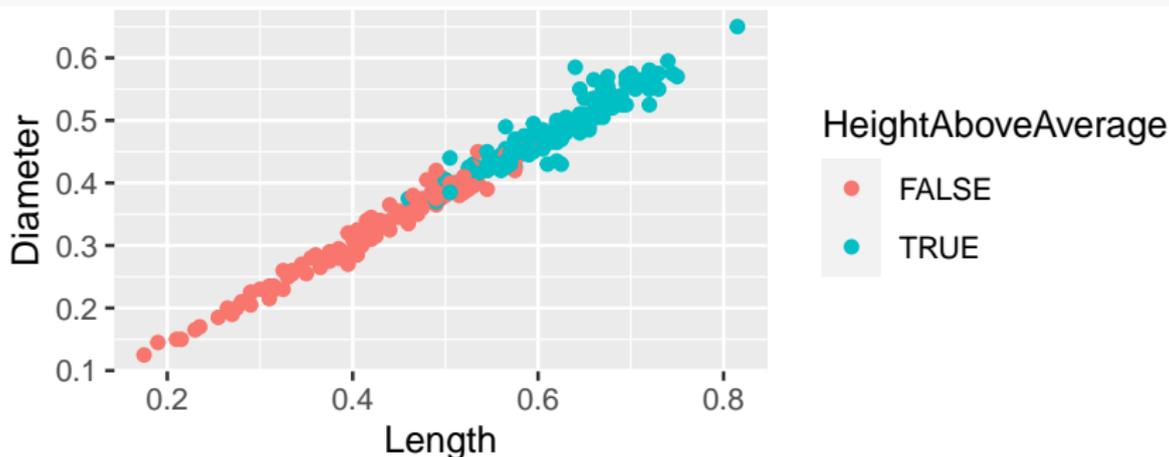
We add a column called `HeightAboveAverage` which will be a logical vector: `TRUE` for all abalones with above average height, `FALSE` for all others. Homework 1, Problem 3 will be helpful for reminding us how to do this:

```
abalone_enhanced <- abalone %>%  
  mutate(HeightAboveAverage=(Height > mean(Height)))  
head(abalone_enhanced$Height - mean(abalone_enhanced$Height))  
## [1] 0.0053 0.0853 0.0003 -0.0297 0.0103 -0.0197  
head(abalone_enhanced$HeightAboveAverage)  
## [1] TRUE TRUE TRUE FALSE TRUE FALSE
```

aes() revisited

Now we color the points based on HeightAboveAverage:

```
abalone_enhanced %>%  
  ggplot(aes(x=Length, y=Diameter)) +  
  geom_point(aes(colour=HeightAboveAverage))
```

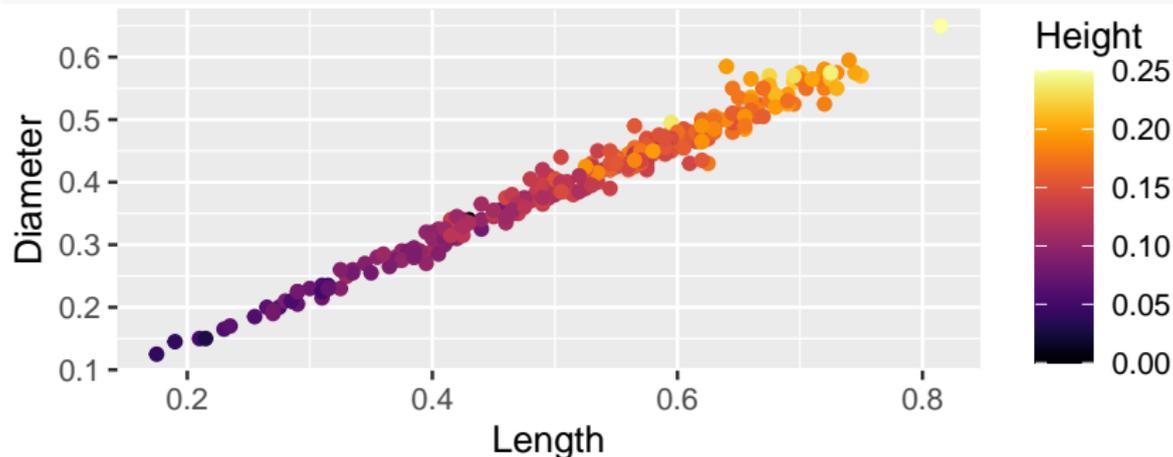


We see that abalones with above average height tend to be longer and have a larger diameter.

aes() revisited

We can also have color vary continuously with height, using a continuous color palette (Lab 5):

```
library(viridis)
## Loading required package: viridisLite
abalone_enhanced %>%
  ggplot(aes(x=Length, y=Diameter)) +
  geom_point(aes(colour=Height)) +
  scale_colour_viridis(option="B")
```



Line plots

Line plots

Line plots are useful for plotting a *quantitative* variable (generally on the vertical axis) against a *sequential* variable (often time, typically on the horizontal axis).

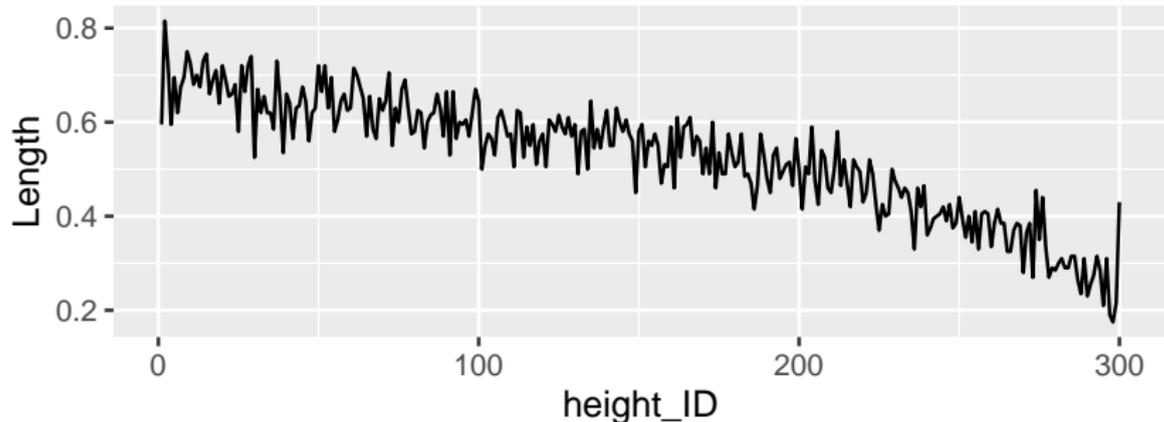
`geom_line()` is the most common geom in `ggplot2` to create line plots. It connects points *in order of the value on the x-axis*.

You could also use `geom_path()` to connect points in the order in which they appear in the rows of the data (see the help pages).

geom_line()

Suppose we assign a `height_ID` to each abalone based on their height order (with 1 being the tallest abalone, 2 being the next tallest, etc.). Let's construct a line plot of `Length` versus `height_ID`:

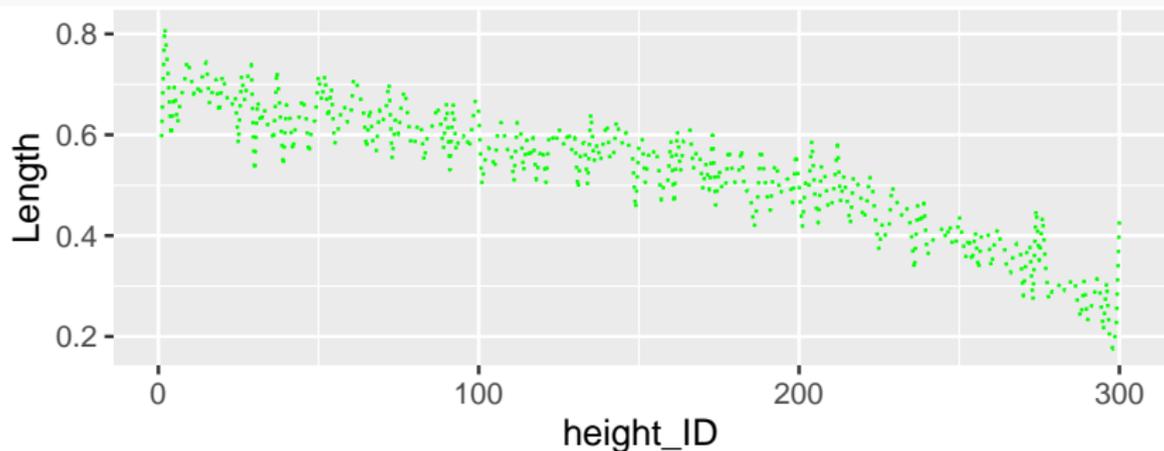
```
abalone %>%  
  arrange(desc(Height)) %>%  
  mutate(height_ID=seq(1, nrow(abalone))) %>%  
  ggplot(aes(x=height_ID, y=Length)) +  
  geom_line()
```



geom_line()

We modify the line type, line thickness, and line color:

```
abalone %>%  
  arrange(desc(Height)) %>%  
  mutate(height_ID=seq(1, nrow(abalone))) %>%  
  ggplot(aes(x=height_ID, y=Length)) +  
  geom_line(linetype="dotted", linewidth=0.5, colour="green")
```



Two categorical variables

Two categorical variables

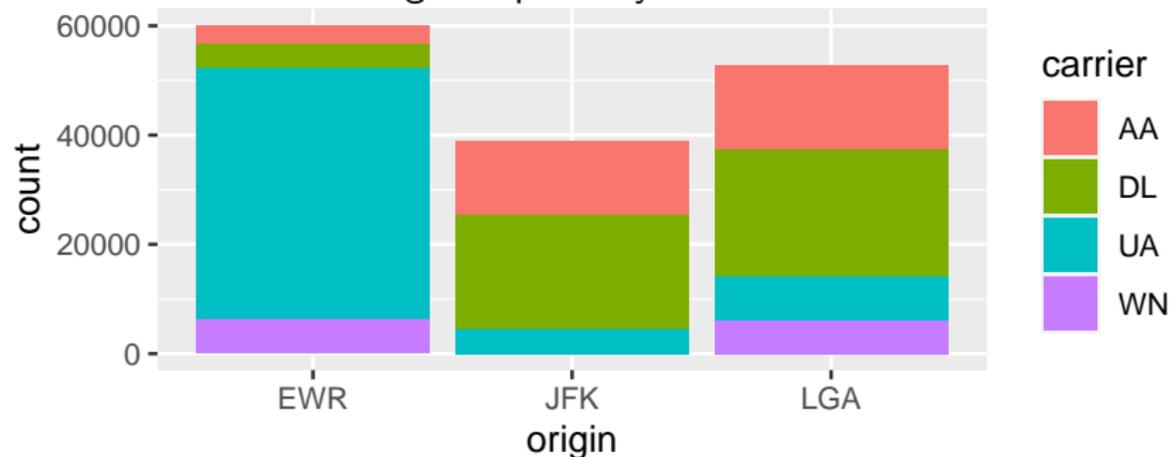
Thus far, we've focused on visualizing the relationship between two quantitative variables.

How do we visualize two categorical variables?

Categorical vs. categorical

A stacked bar chart is typically the best way to view categorical vs. categorical data.

In fact, we've already used these, when demonstrating `aes()` in the last lecture to show origin airports by carrier:

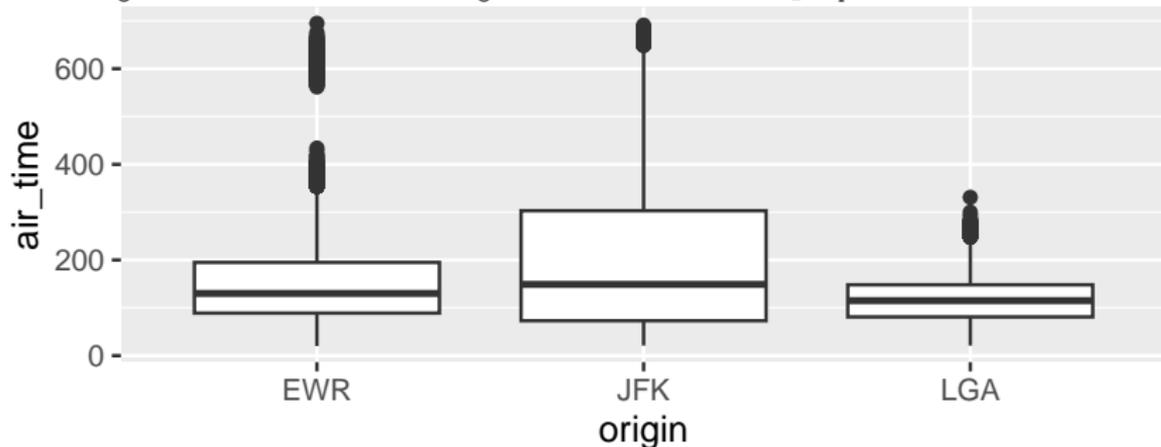


Quantitative vs. categorical

Side-by-side boxplots are a common choice to visualize one quantitative vs. one categorical variable. They are created by specifying both an x and y aesthetic for `geom_boxplot()`:

```
flights %>%  
  ggplot(aes(x=origin, y=air_time)) +  
  geom_boxplot()
```

```
## Warning: Removed 9430 rows containing non-finite values (`stat_boxplot()`).
```



More than two variables

3-D plots are typically hard to read on a flat screen or sheet of paper.

Instead, you should use aesthetics when possible.

Summary of main plot types

Summary of main plot types

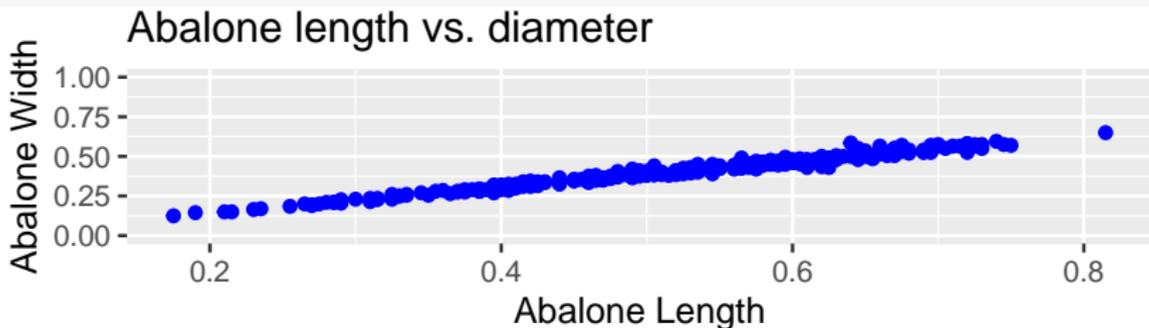
Plot Type	When to Use
Scatterplot	Compare two quantitative variables
Line plot	Plot one quantitative variable against a sequential variable
(Stacked) bar plot	Understand distribution of one (or more) categorical variable(s)
(Side-by-side) Boxplot	Understand distribution of one quantitative variable (based on another categorical variable)
Histogram	Understand distribution of one quantitative variable

Layers

Layers

Titles, axis limits, and axis labels are examples of additional **layers** you can add to your plot. Geoms are also layers; indeed, anything added on to the `ggplot()` call with `+` is a layer. Let's add a title, axis labels, and y axis limits:

```
abalone %>%  
  ggplot(aes(x=Length, y=Diameter)) +  
  geom_point(colour="blue") +  
  ggtitle("Abalone length vs. diameter") +  
  xlab("Abalone Length") +  
  ylab("Abalone Width") +  
  ylim(c(0, 1))
```



Faceting

Faceting

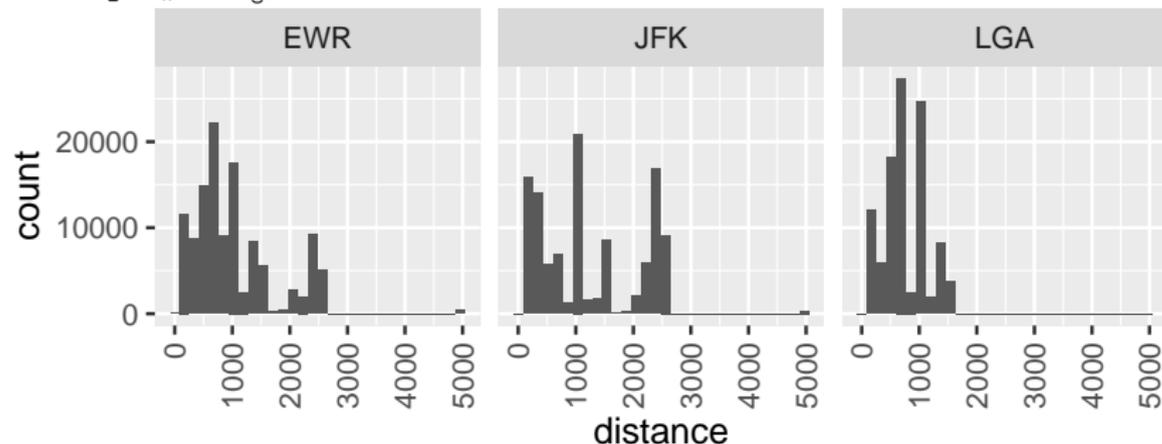
Faceting occurs when you want to see multiple plots broken down by some categorical variable.

We use `facet_wrap()` to visualize the distribution of flight distances for each `origin`.

Faceting

```
flights %>%  
  ggplot(aes(x=distance)) +  
  geom_histogram() +  
  facet_wrap(~origin) +  
  theme(axis.text.x=element_text(angle=90, vjust=0.5, hjust=1))
```

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



Faceting

Flights from LGA tend to be the shortest (perimeter rule).

JFK has more long flights (over 2000 miles) but fewer shorter flights (500-1000 miles) than EWR.

Faceted plots provide a lot more info than summary statistics. A picture is worth 1,000 words!