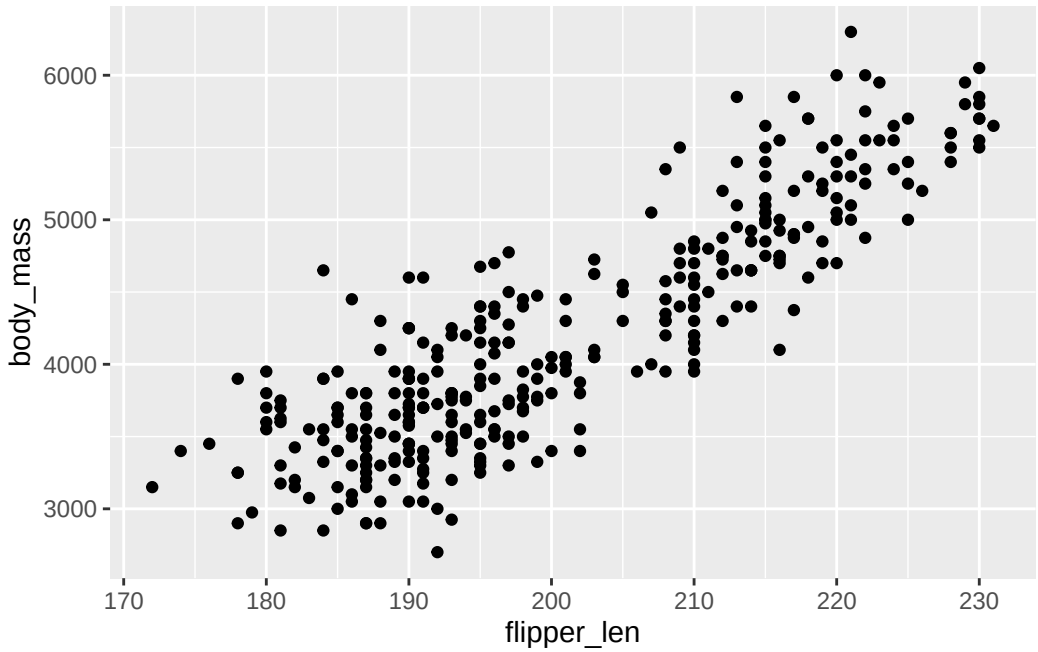
R Basics

Aheer Srabon

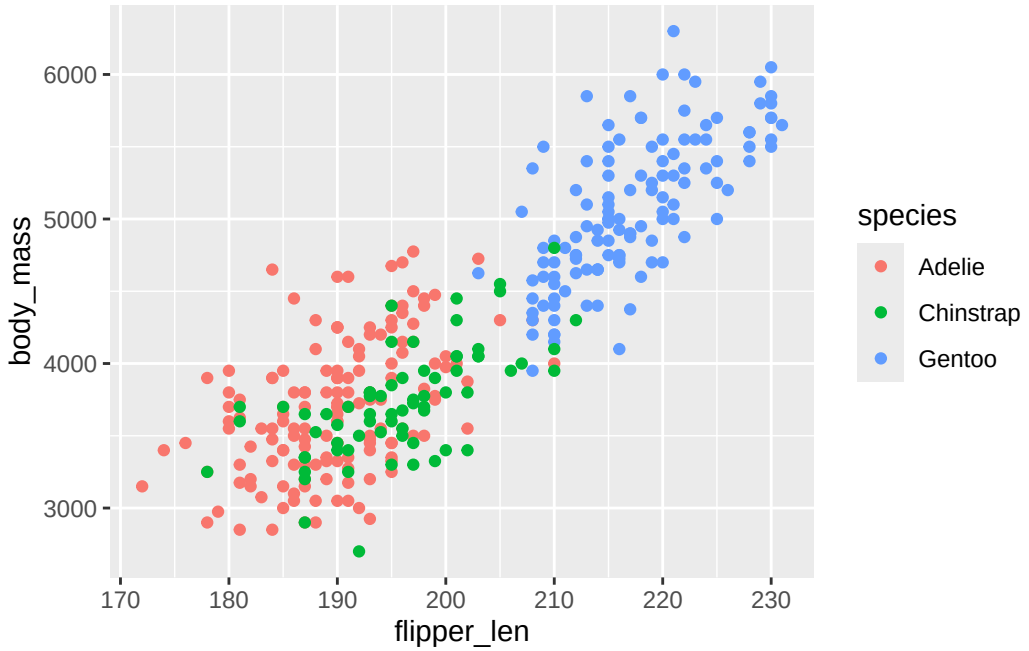
Visualizing distributions

```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_len, y = body_mass)
) +
  geom_point()
```

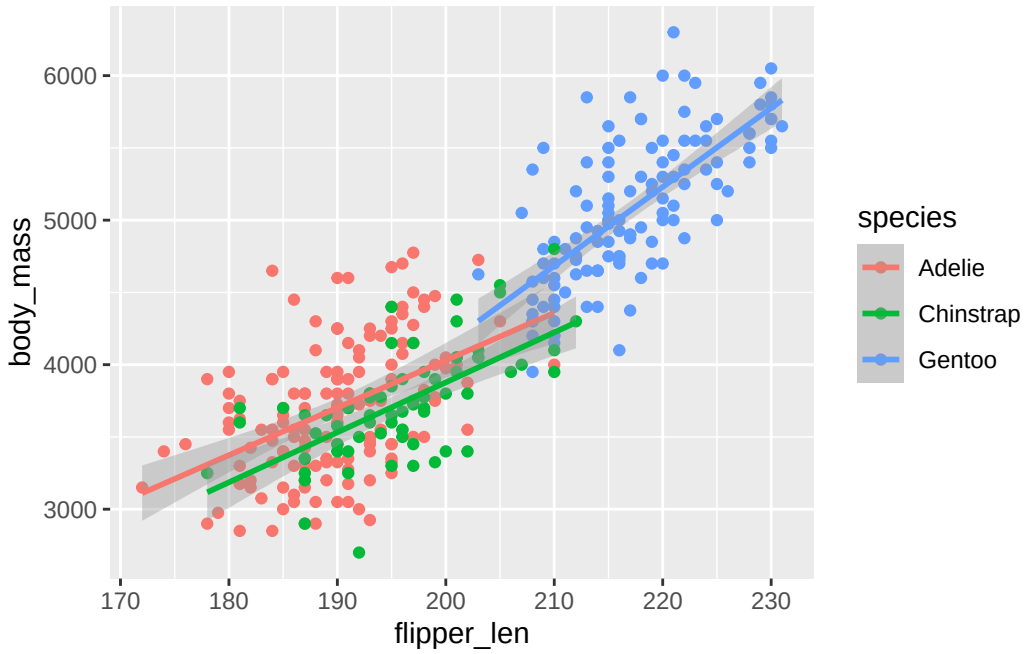



```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_len, y = body_mass, color = species)
```

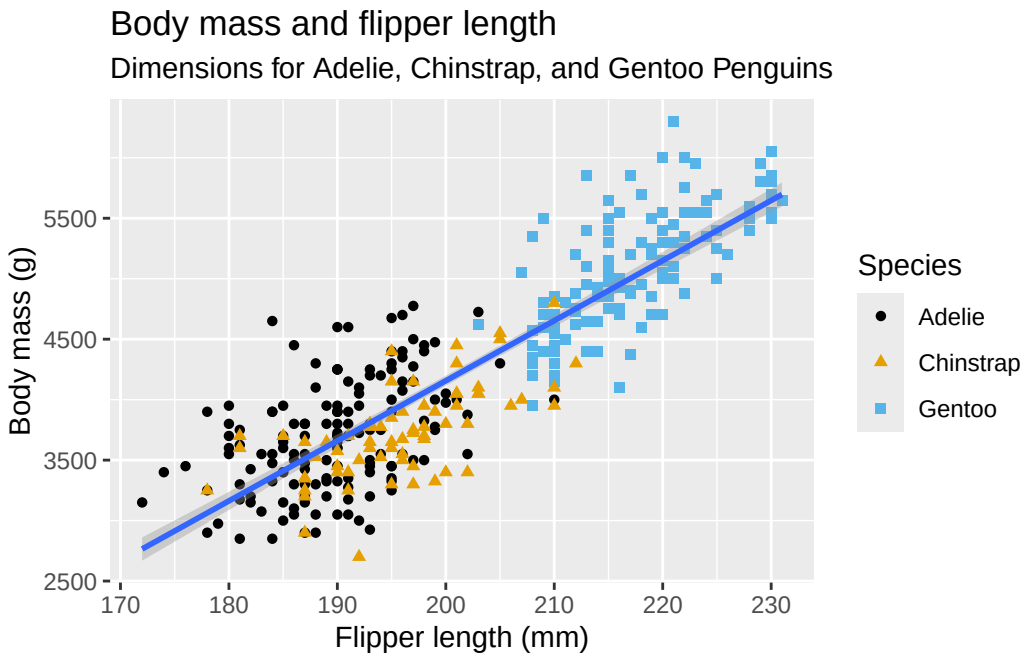
```
) +
  geom_point()
```



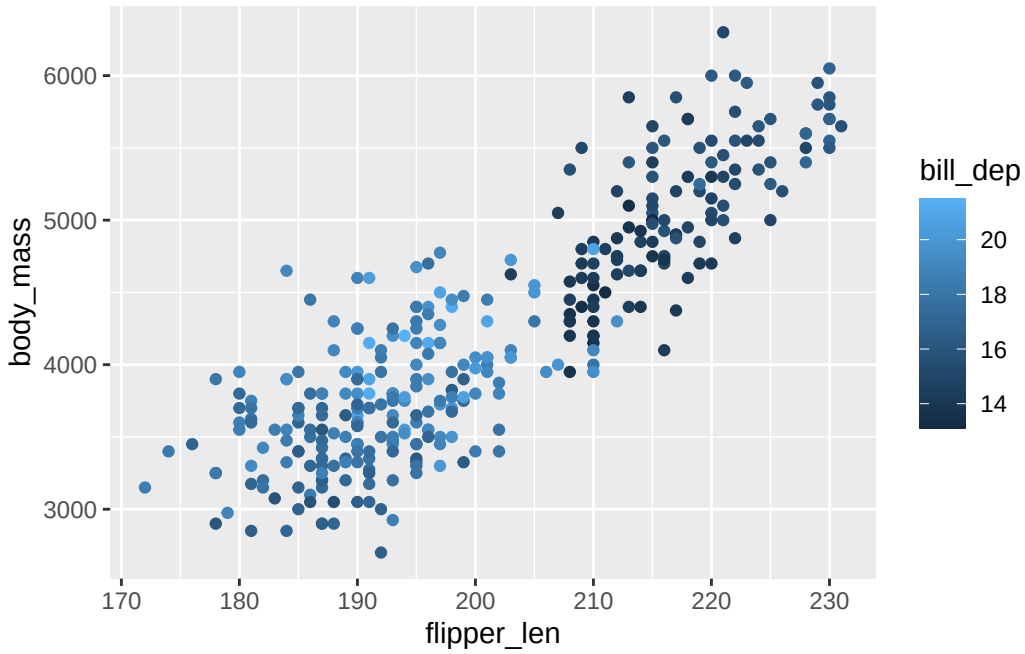

```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_len, y = body_mass, color = species)
) +
  geom_point() +
  geom_smooth(method = "lm")
```



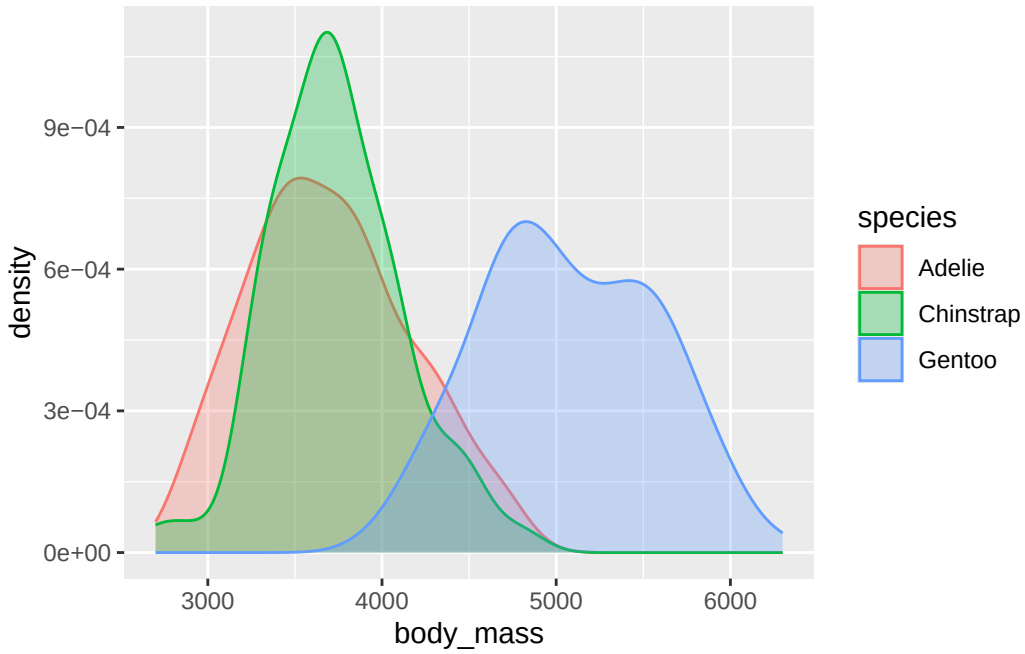

```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_len, y = body_mass)
) +
  geom_point(mapping = aes(color = species, shape = species)) +
  geom_smooth(method = "lm") +
  labs(
    title = "Body mass and flipper length",
    subtitle = "Dimensions for Adelie, Chinstrap, and Gentoo Penguins",
    x = "Flipper length (mm)", y = "Body mass (g)",
    color = "Species", shape = "Species"
  ) +
  scale_color_colorblind()
```



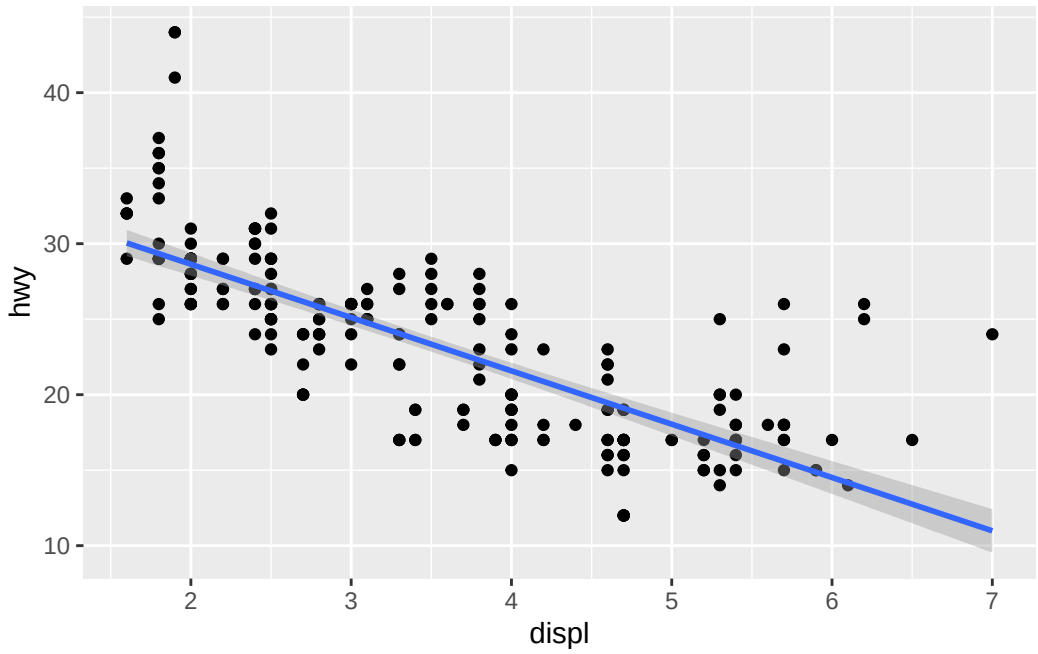

```
ggplot(data = penguins, mapping = aes(
  x = flipper_len,
  y = body_mass
)) +
  geom_point(mapping = aes(
    color = bill_dep))
```




```
ggplot(penguins, aes(
  x = body_mass,
  color = species,
  fill = species
)) +
  geom_density(alpha = 0.3)
```




```
ggplot(mpg, mapping = aes(
  x = displ,
  y = hwy
)) +
  geom_point() +
  geom_smooth(method = "lm")
```



Data transformation

```
# had an arrival delay of more than two hours
flights |>
  filter(arr_delay > 2) |>
  arrange(arr_delay)
```

```
# A tibble: 123,096 x 19
    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
 1  2013     1     1      622            630        -8     1017           1014
 2  2013     1     1      728            732        -4     1041           1038
 3  2013     1     1      743            730        13     1059           1056
 4  2013     1     1      830            830         0     1018           1015
 5  2013     1     1      902            903        -1     1048           1045
 6  2013     1     1      937            940        -3     1238           1235
 7  2013     1     1     1113           1115        -2     1318           1315
 8  2013     1     1     1130           1131        -1     1345           1342
 9  2013     1     1     1133           1129         4     1440           1437
10  2013     1     1     1231           1238        -7     1449           1446
# i 123,086 more rows
```

```
# i 11 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
#   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
#   hour <dbl>, minute <dbl>, time_hour <dttm>
```

```
# flew to Houston
flights |>
  filter(dest == "IAH" | dest == "HOU") |>
  arrange(dest)
```

```
# A tibble: 9,313 x 19
    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
 1  2013     1     1     1208           1158        10     1540           1502
 2  2013     1     1     1306           1300         6     1622           1610
 3  2013     1     1     1708           1700         8     2037           2005
 4  2013     1     1     2030           2035        -5     2354           2342
 5  2013     1     2      734            700        34     1045           1025
 6  2013     1     2     1156           1158        -2     1517           1502
 7  2013     1     2     1319           1305        14     1633           1615
 8  2013     1     2     1810           1655        75     2146           2000
 9  2013     1     2     2031           2035        -4     2353           2342
10  2013     1     3      704            700         4     1036           1025
# i 9,303 more rows
# i 11 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
#   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
#   hour <dbl>, minute <dbl>, time_hour <dttm>
```

```
# were operated by United, American, or Delta
flights |>
  filter(carrier == "UA" | carrier == "AA" | carrier == "DL") |>
  arrange(carrier)
```

```
# A tibble: 139,504 x 19
    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
 1  2013     1     1      542            540         2      923            850
 2  2013     1     1      558            600        -2      753            745
 3  2013     1     1      559            600        -1      941            910
 4  2013     1     1      606            610        -4      858            910
 5  2013     1     1      623            610        13      920            915
 6  2013     1     1      628            630        -2     1137           1140
```

```
 7  2013     1     1      629            630        -1      824            810
 8  2013     1     1      635            635         0     1028            940
 9  2013     1     1      656            700        -4      854            850
10  2013     1     1      656            659        -3      949            959
# i 139,494 more rows
# i 11 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
#   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
#   hour <dbl>, minute <dbl>, time_hour <dttm>
```

```
# departed in summar (july, august, or september)
flights |>
  filter(month >=7 & month <= 9) |>
  arrange(month)
```

```
# A tibble: 86,326 x 19
    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
 1  2013     7     1        1           2029       212      236           2359
 2  2013     7     1        2           2359         3      344            344
 3  2013     7     1       29           2245       104      151              1
 4  2013     7     1       43           2130       193      322             14
 5  2013     7     1       44           2150       174      300            100
 6  2013     7     1       46           2051       235      304           2358
 7  2013     7     1       48           2001       287      308           2305
 8  2013     7     1       58           2155       183      335             43
 9  2013     7     1      100           2146       194      327             30
10  2013     7     1      100           2245       135      337            135
# i 86,316 more rows
# i 11 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
#   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
#   hour <dbl>, minute <dbl>, time_hour <dttm>
```

```
# Arrived more than two hours late but didn't leave late
flights |>
  filter(arr_delay > 2 & dep_delay == 0) |>
  arrange(arr_delay)
```

```
# A tibble: 4,368 x 19
    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
 1  2013     1     1      830            830         0     1018           1015
```

```
 2  2013     1     1     2040           2040           0     2317           2314
 3  2013     1     3     1225           1225           0     1351           1348
 4  2013     1     8      700            700           0      810            807
 5  2013     1    11      700            700           0     1000            957
 6  2013     1    11     1810           1810           0     2145           2142
 7  2013     1    11     2025           2025           0     2332           2329
 8  2013     1    13      928            928           0     1054           1051
 9  2013     1    14     1545           1545           0     1820           1817
10  2013     1    15     1301           1301           0     1407           1404
# i 4,358 more rows
# i 11 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
#   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
#   hour <dbl>, minute <dbl>, time_hour <dttm>
```

```
# Were delayed by at least an hour, but made up over 30 minutes in flight
flights |>
  filter(dep_delay >= 1 & air_time > 30) |>
  select(air_time, dep_delay, origin, dest) |>
  arrange(air_time)
```

```
# A tibble: 127,205 x 4
   air_time dep_delay origin dest
      <dbl>     <dbl> <chr>  <chr>
 1       31        52 EWR    ALB
 2       31        85 EWR    ALB
 3       31        57 EWR    ALB
 4       31        31 EWR    ALB
 5       31        15 JFK    BOS
 6       31        74 EWR    PHL
 7       31        36 JFK    PHL
 8       31        62 EWR    PHL
 9       31        10 EWR    BOS
10       31         8 JFK    BOS
# i 127,195 more rows
```

```
# Sort flights to find the flights with the longest departure delays.
# Find the flights that left earliest in the morning.
```