Introduction to R (TAU Workshop) Visualizations

Telling stories with charts (Third session)

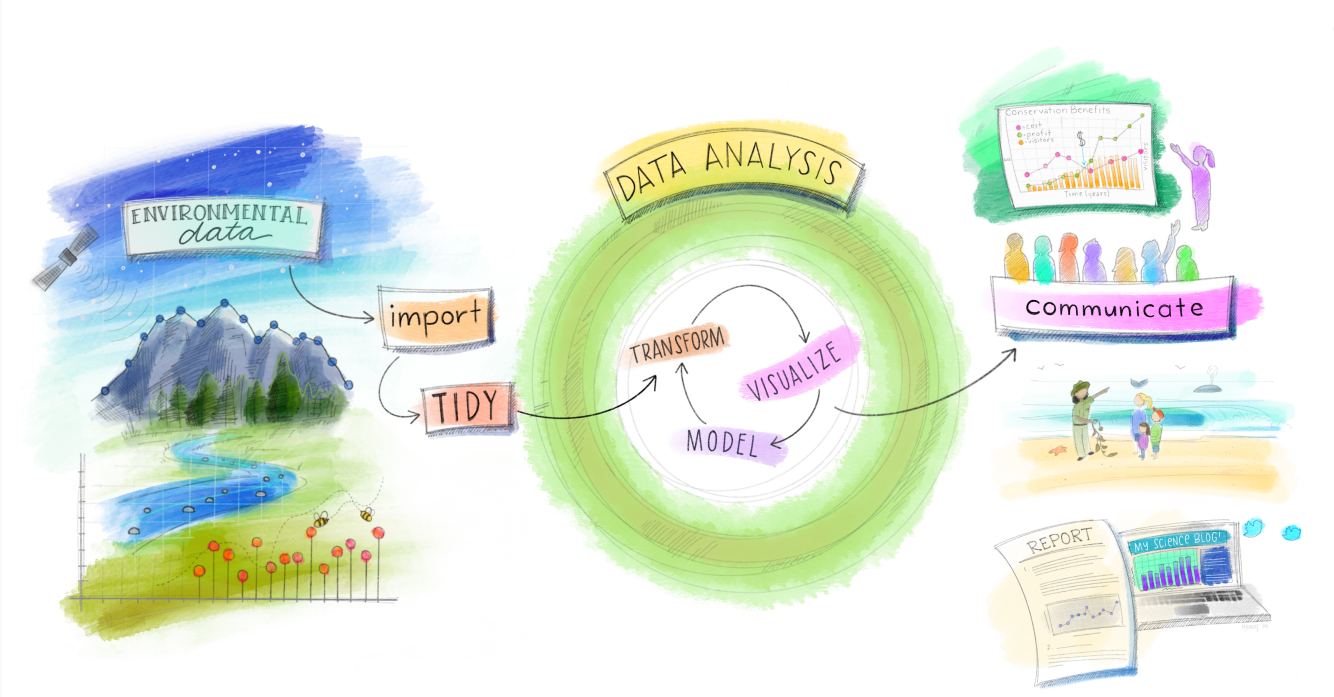
Adi Sarid
Tel-Aviv University
Updated: 2020-12-21

Recap from last week

In the previous session

- We have seen a lot of tidyverse basics for data transformations and wrangling
  - Rows affecting functions (e.g., `filter`), column affecting functions (e.g., `mutate`)
  - We saw an aggregating function (`count`), and data introduction function (`glimpse`)
  - You want to exercise these as much as you can if you want to become proficient in data science
- You experienced the various functions using the Himalayan Climbing Expeditions data set from the tidytuesday repository

Today's focus: data wrangling continued, and visualizations



What we will do today

- Complete some things from last week
  - Operations that summarize data (e.g., compute mean, sd, percentiles)
- Discuss the grammar of graphics
- Learn and exercise `ggplot2`

Operations that summarize data (summarize cases)

Apply summary functions to create a new table. To summarize data you need to combine `group_by`, `summarize`, and a summary function:

- `group_by` and `summarize`, followed by:
- `sum`, `n`, `mean`, `median`, `first`, `last`, `quantile`, `min`, `max`, `sd`

(Examples: line 108 in `02-Data-Transformations-and-Wrangling.R`)

The datasarus dozen

Why are visualizations important?

- It is an amazing way to communicate findings
- Sometimes it's the only way

A classic example in the "Datasarus Dozen dataset"

```
library(datasauRus)
glimpse(datasaurus_dozen)
```

```
## Rows: 1,846
## Columns: 3
## $ dataset <chr> "dino", "dino", "dino", "dino", "dino", "dino", "dino", "di...
## $ x       <dbl> 55.3846, 51.5385, 46.1538, 42.8205, 40.7692, 38.7179, 35.64...
## $ y       <dbl> 97.1795, 96.0256, 94.4872, 91.4103, 88.3333, 84.8718, 79.87...
```

The datasarus dozen (mean, sd, and correlation)

What is the relationship between `x` and `y`?

We can group the datasets within datasarus_dozen and compute mean, sd, and correlation.

```
datasaurus_dozen %>%
   group_by(dataset) %>%
   summarize(
      mean_x = mean(x),
      mean_y    = mean(y),
      std_dev_x = sd(x),
      std_dev_y = sd(y),
      corr_x_y  = cor(x, y)
   )
```

The datasarus dozen (mean, sd, and correlation)

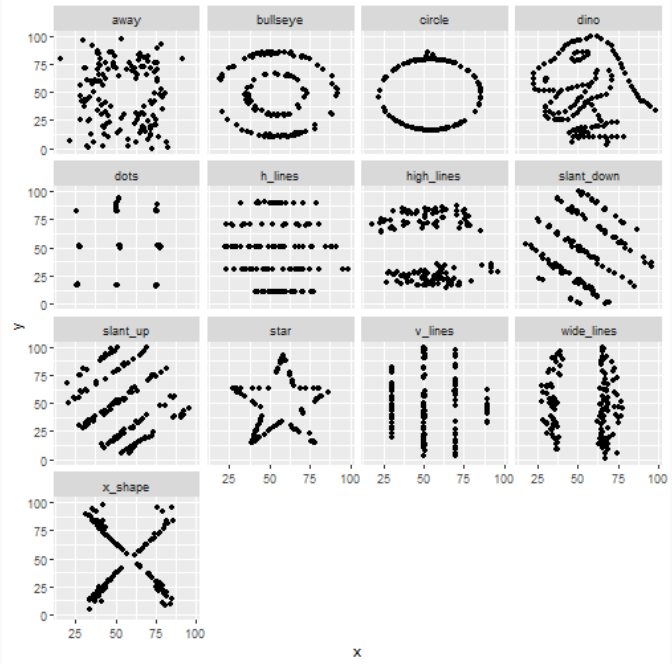
What is the relationship between `x` and `y`?

```
## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 13 x 6
##    dataset    mean_x mean_y std_dev_x std_dev_y corr_x_y
##    <chr>       <dbl>  <dbl>     <dbl>     <dbl>    <dbl>
##  1 away         54.3   47.8      16.8      26.9  -0.0641
##  2 bullseye     54.3   47.8      16.8      26.9  -0.0686
##  3 circle       54.3   47.8      16.8      26.9  -0.0683
##  4 dino         54.3   47.8      16.8      26.9  -0.0645
##  5 dots         54.3   47.8      16.8      26.9  -0.0603
##  6 h_lines      54.3   47.8      16.8      26.9  -0.0617
##  7 high_lines   54.3   47.8      16.8      26.9  -0.0685
##  8 slant_down   54.3   47.8      16.8      26.9  -0.0690
##  9 slant_up     54.3   47.8      16.8      26.9  -0.0686
## 10 star         54.3   47.8      16.8      26.9  -0.0630
## 11 v_lines      54.3   47.8      16.8      26.9  -0.0694
## 12 wide_lines   54.3   47.8      16.8      26.9  -0.0666
## 13 x_shape      54.3   47.8      16.8      26.9  -0.0656
```

The datasarus dozen - IT'S MAGIC! (Simpson's paradox)

```
ggplot(datasaurus_dozen) + geom_point(aes(x, y)) + facet_wrap(~dataset)
```



The grammar of graphics

Each variable in the data is mapped by an "aesthetic mapping":

- Axis (x, y); Color; Fill; Alpha (transparency); Size; Shape; etc.

The aesthetics are combined with geometric functions, e.g.:

- Points, bars, lines, histograms, densities, etc (there are a lot of them)

These elements are implemented in the `ggplot2` package (which you've seen in the first session of this workshop)

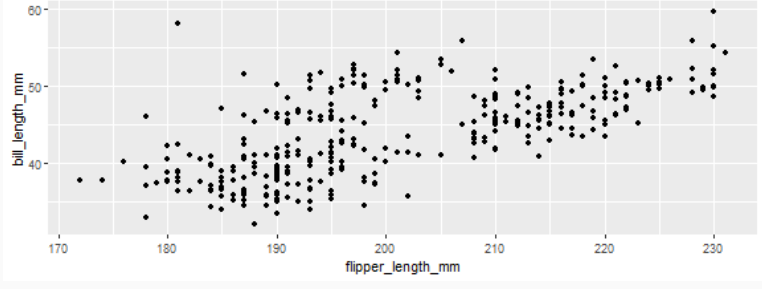
- We have a template! (and a cheat sheet)

```
ggplot(data = <DATA>, mapping = aes(<GLOBAL MAPPINGS>)) +
   <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```

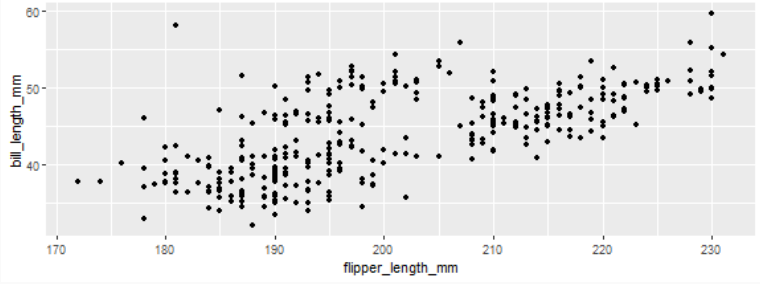
Example: relationship of penguin flipper and bill length

```
library(palmerpenguins)
ggplot(penguins) +
   geom_point(aes(x = flipper_length_mm, y = bill_length_mm))
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



Example: relationship of penguin flipper and bill length (questions)

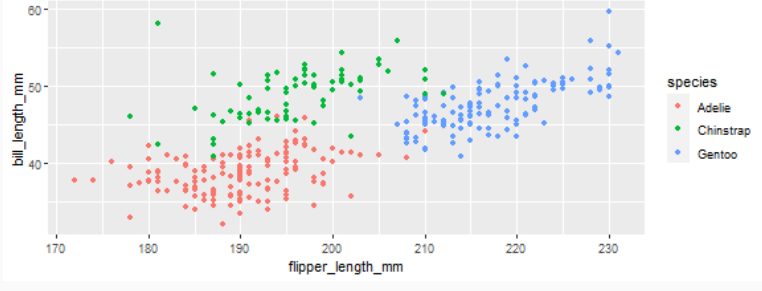


- What story does this chart tells you?
- Is there a difference between penguin species?
- What is the aesthetic mapping?
- Why is there a warning message?

Species matters

```
ggplot(penguins) +
   geom_point(aes(x = flipper_length_mm, y = bill_length_mm,
                  color = species))
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

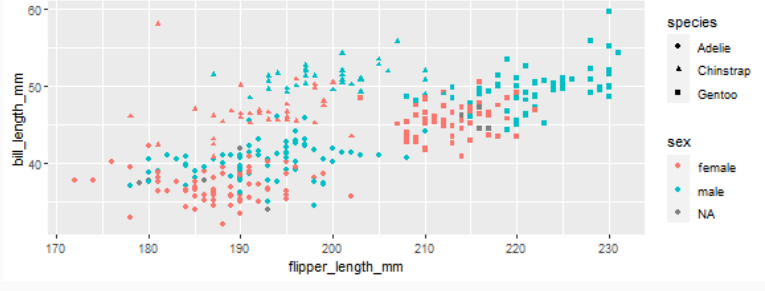


- What did you learn now that you didn't know before?
- If you wanted to add gender (`sex`) to this comparison, how would you do that?

Gender matters as well

```
ggplot(penguins) +
    geom_point(aes(x = flipper_length_mm, y = bill_length_mm,
                   shape = species, color = sex))
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

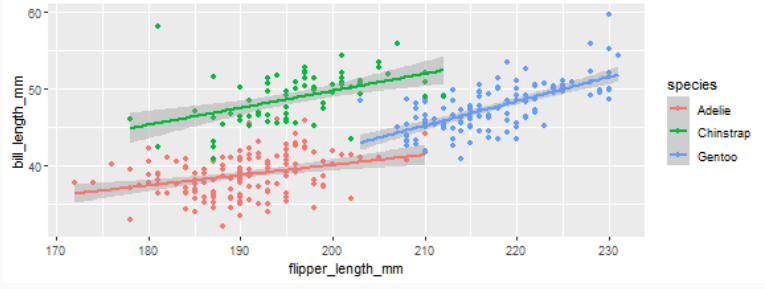


- What information was added to this chart?
- The gender missing values, can you guess what they are?

Adding regression models

```
ggplot(penguins, aes(x = flipper_length_mm, y = bill_length_mm, color = species)) +
   geom_point() + geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



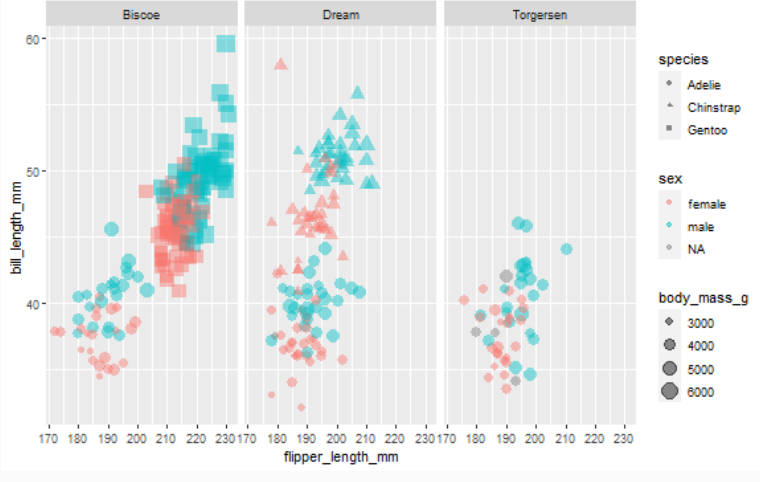
- Note the use of global aesthetic mappings
- Note that you can add layers one after the other (a number of geoms)

WARNING: Don't get carried away

There is such as a thing as too many details.

- The human mind can process 7 ± 2 items in short term memory (see here)

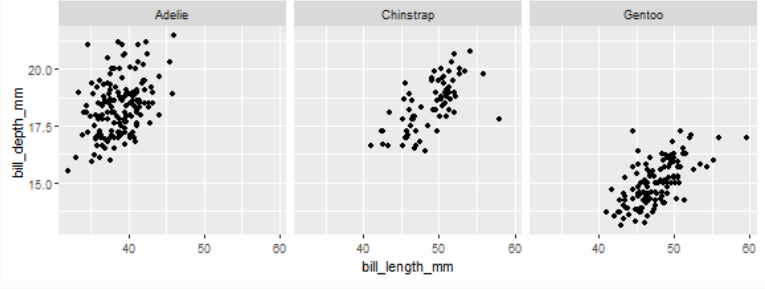
```
ggplot(penguins) + geom_point(aes(x = flipper_length_mm, y = bill_length_mm, shape = species, color = sex, size = body_mass_g), alpha = 0.45) + facet_wrap(~islar
```



Facets

Sometimes you need to "split" the plots according to a specific variable in order to make a point. This is called faceting. For example, we can replace the aesthetic mapping of `color=species` with faceting

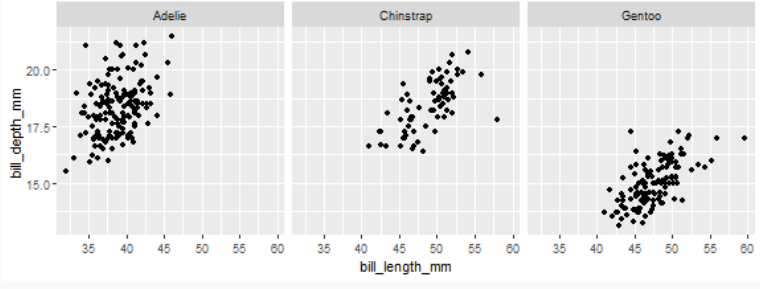
```
ggplot(penguins) +
   geom_point(aes(x = bill_length_mm, y = bill_depth_mm)) +
   facet_wrap(~species)
```



Scales

Sometimes, we want to slightly modify aesthetic mappings. This can be accomplished using scales.

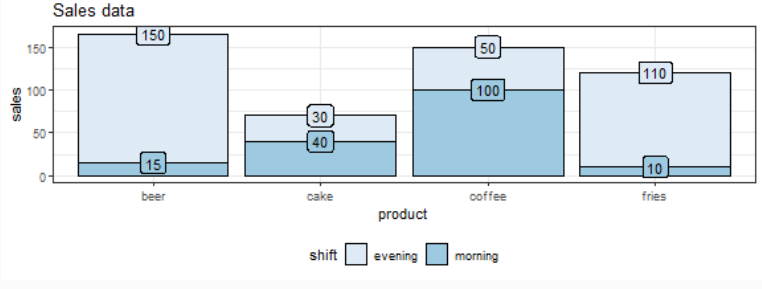
```
ggplot(penguins) +
   geom_point(aes(x = bill_length_mm, y = bill_depth_mm)) + facet_wrap(~species) +
   scale_x_continuous(breaks = seq(30, 60, by = 5))
```



We will see some more interesting examples (i.e., log scales) in the following sessions.

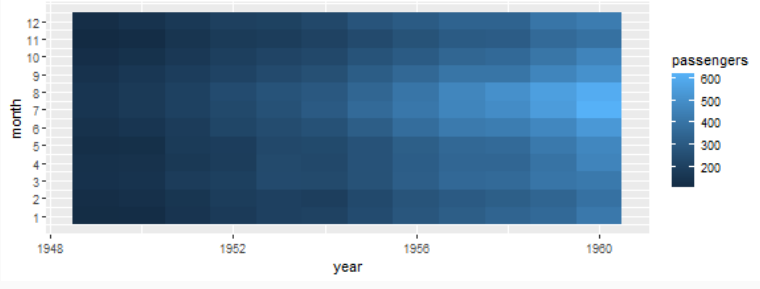
Mini exercise: how would you...? (1)

- Using the `ggplot2` cheat sheet, what are the three geoms required to produce this chart?
- What are the aesthetic mappings?



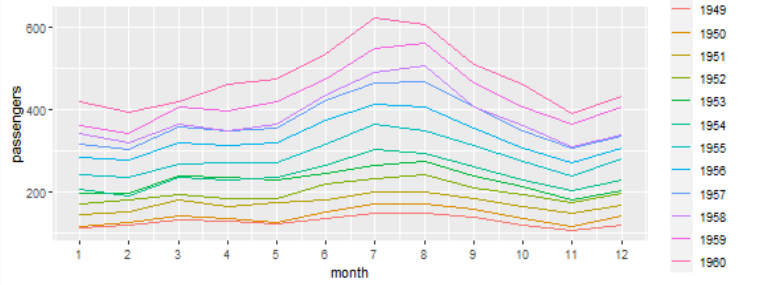
Mini exercise: how would you...? (2)

- Monthly Airline passenger numbers 1949-1960 (see `?AirPassengers`)
- Use the ggplot2 cheat sheet
- What is the **one** geom required to produce this chart?
- What are the aesthetic mappings?



Mini exercise: how would you...? (3)

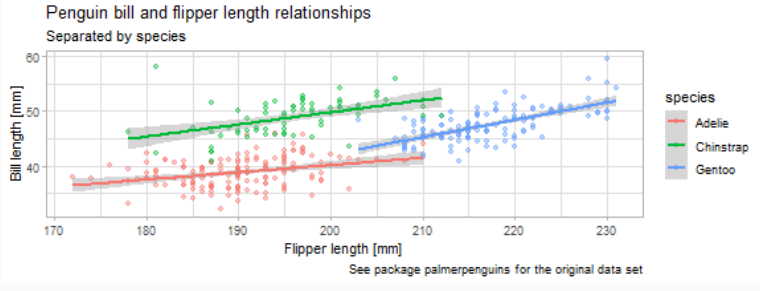
Same passengers data, different visualization: what are the aesthetics and geom? which visualization is better and why?



Final touch ups

You can control every aspect of the chart. For example, you can change the theme, the labels and the title.

```
ggplot(penguins, aes(x = flipper_length_mm, y = bill_length_mm, color = species)) +
   geom_point(alpha = 0.4) + geom_smooth(method = "lm") +
   labs(title = "Penguin bill and flipper length relationships",
        subtitle = "Separated by species",
        caption = "See package palmerpenguins for the original data set") +
   xlab("Flipper length [mm]") +
   ylab("Bill length [mm]") +
   theme_light()
```



Exercise

In `labs/Tidying Himalayan Climbing Expeditions/03-Visualization-Himalayan-Exercise.R`, also available here.

In this exercise you will continue your work on the Himalayan climbing expeditions data (which we started last week).

Wrap up

- We discussed operations that summarize data (e.g., compute mean, sd, percentiles)
- We talked about the grammar of graphics
- We demonstrated and exercised ggplot2
  - Aesthetic mappings
  - Geoms
  - Scales
  - Facets