# Code Along

October 1st, 2024

```
library(palmerpenguins)
library(ggplot2)
library(dplyr)

data("penguins")

summary(penguins)
```

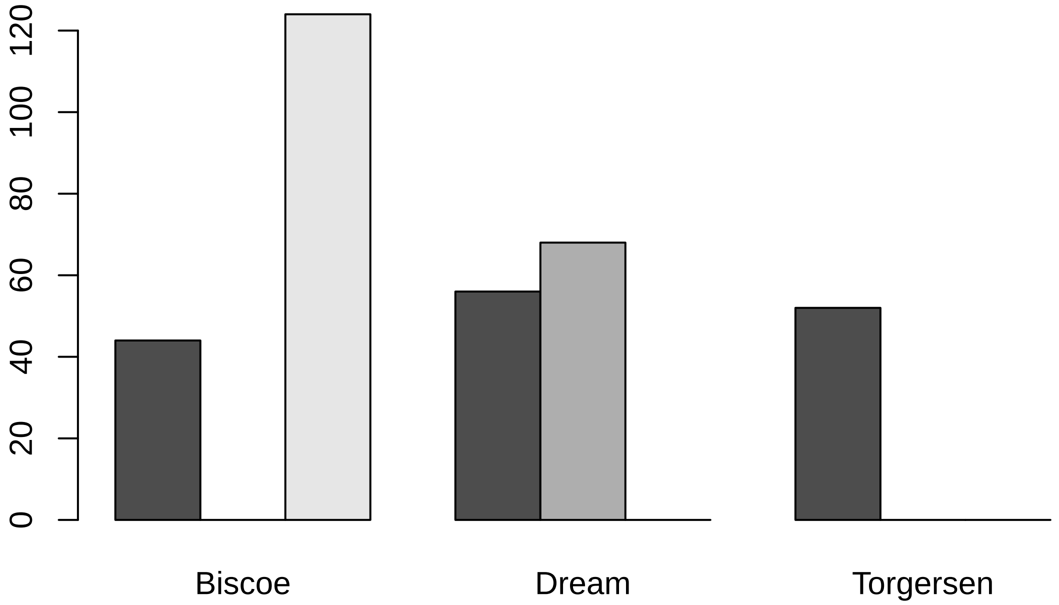
```
##       species          island    bill_length_mm  bill_depth_mm  
##  Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10  
##  Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60  
##  Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30  
##                                  Mean   :43.92   Mean   :17.15  
##                                  3rd Qu.:48.50   3rd Qu.:18.70  
##                                  Max.   :59.60   Max.   :21.50  
##                                  NA's   :2       NA's   :2      
##  flipper_length_mm  body_mass_g       sex           year     
##  Min.   :172.0     Min.   :2700   female:165   Min.   :2007  
##  1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007  
##  Median :197.0     Median :4050   NA's  : 11   Median :2008  
##  Mean   :200.9     Mean   :4202                Mean   :2008  
##  3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009  
##  Max.   :231.0     Max.   :6300                Max.   :2009  
##  NA's   :2         NA's   :2
```

### EDA

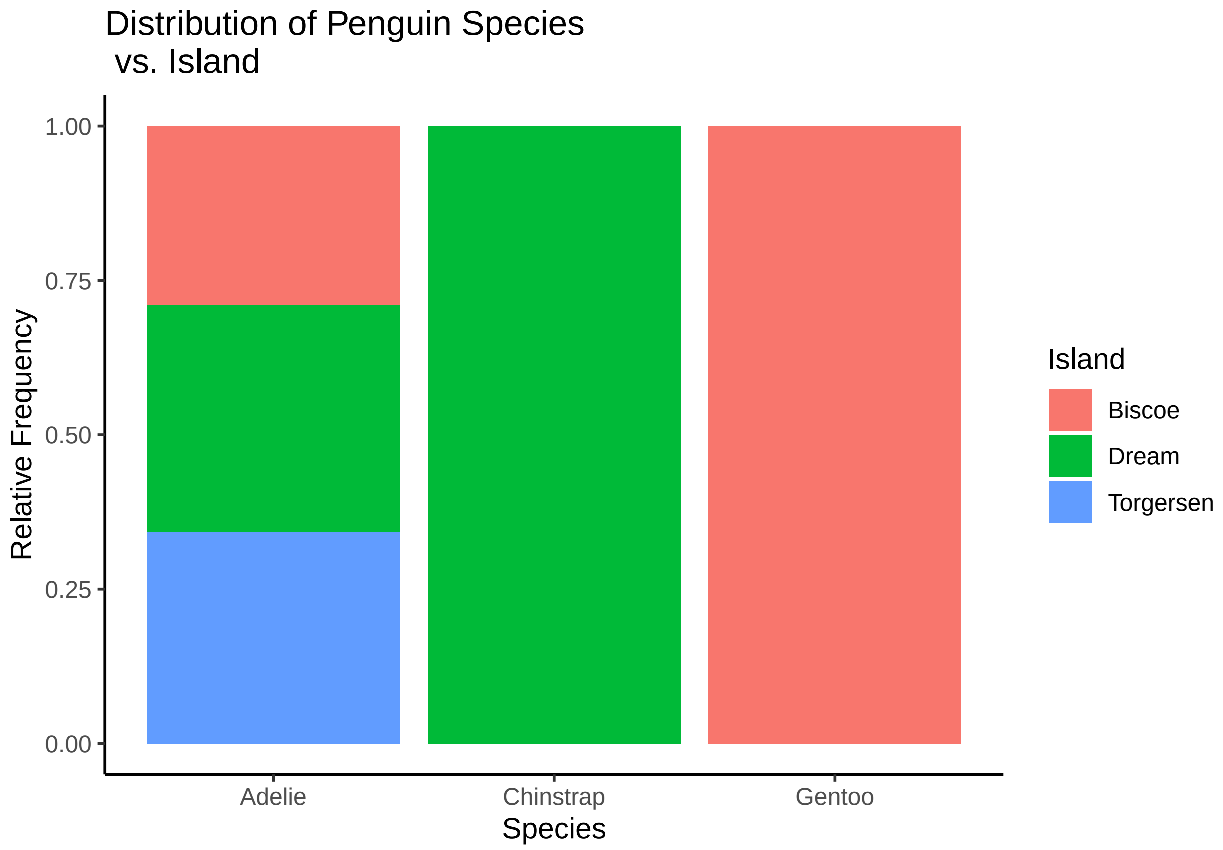
```
table(penguins$species, penguins$island)
```

```
## 
##             Biscoe Dream Torgersen
##   Adelie        44    56        52
##   Chinstrap      0    68         0
##   Gentoo       124     0         0
```

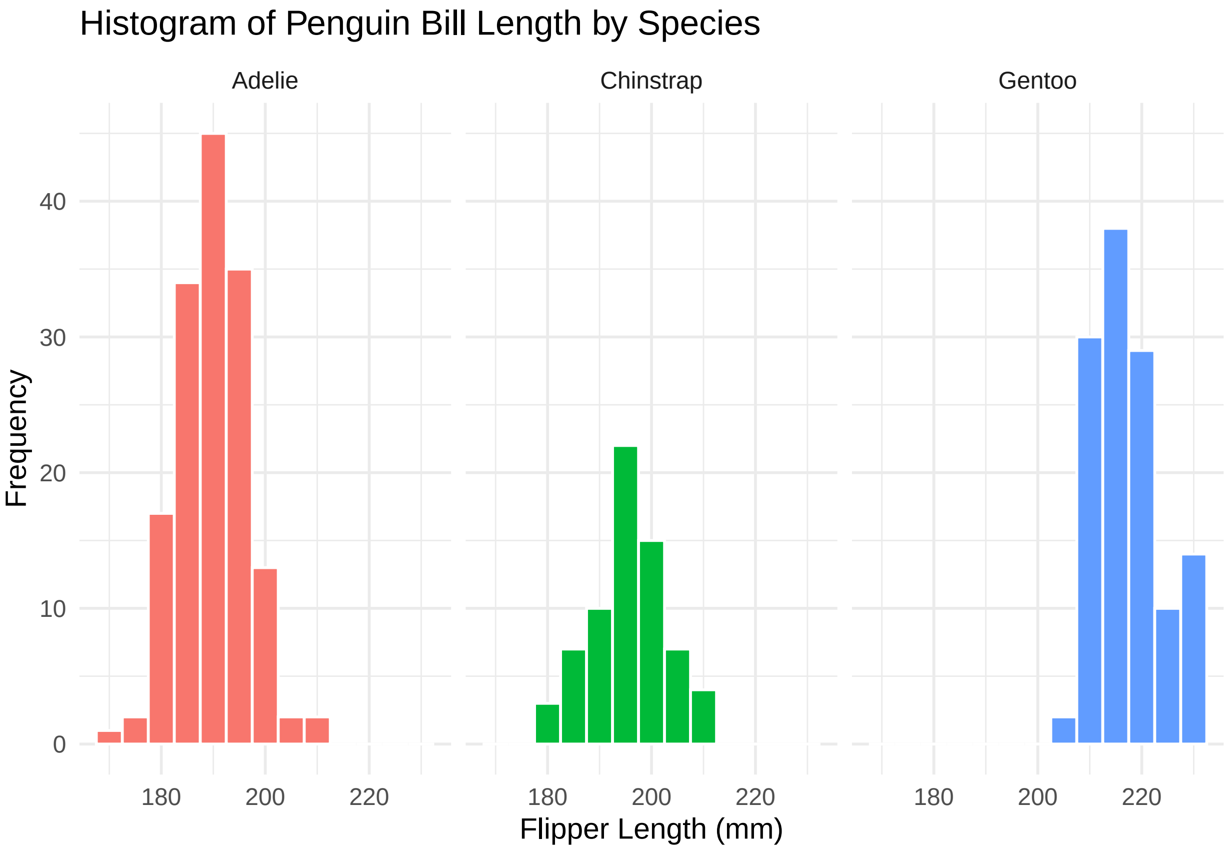
```
barplot(table(penguins$species, penguins$island),
        beside = TRUE)
```



```
ggplot(penguins, aes(x = species,
                     fill = island)) +
  geom_bar(position = "fill") +
  labs(title = "Distribution of Penguin Species\n vs. Island",
       x = "Species",
       y = "Relative Frequency",
       fill = "Island") +
  theme_classic()
```



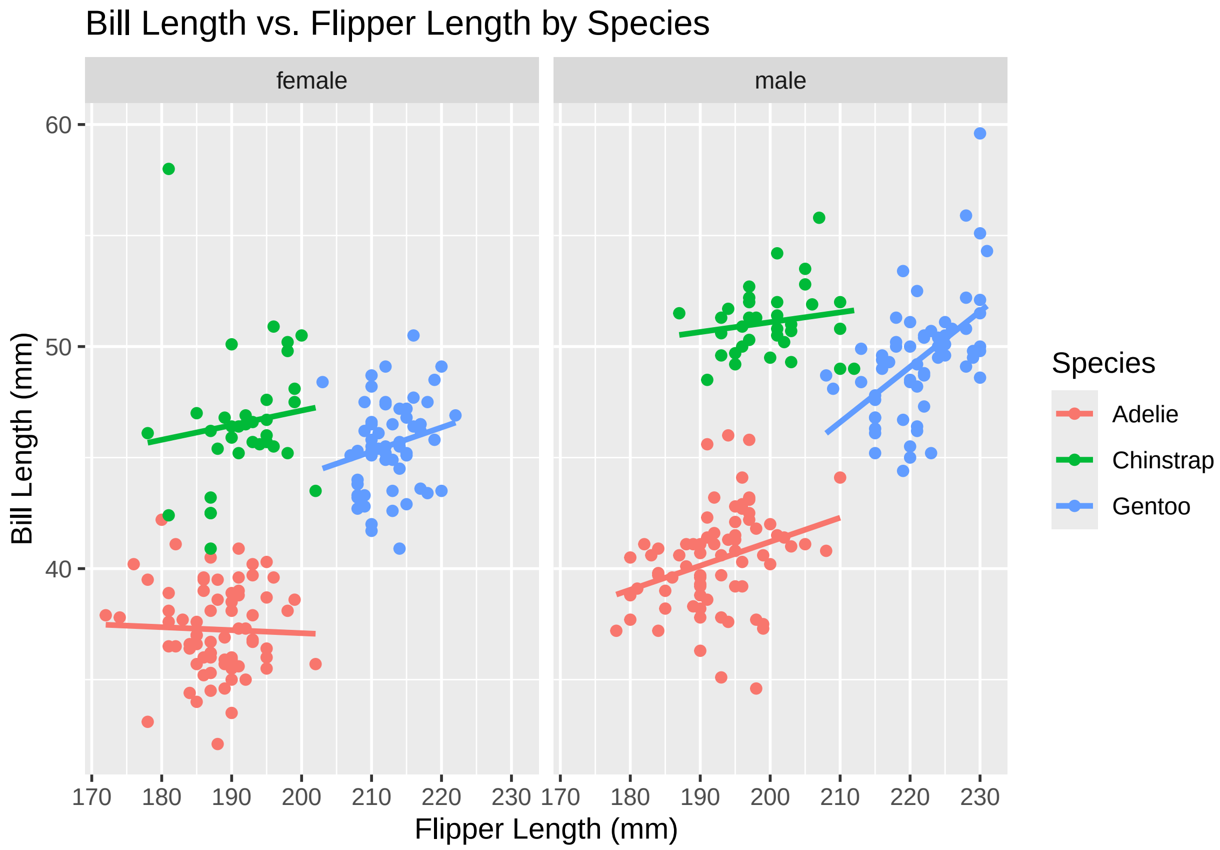
```
ggplot(penguins, aes(x = flipper_length_mm,
                     fill = species)) +
  geom_histogram(binwidth = 5,
                 col = "white",
                 show.legend = FALSE) +
  facet_wrap(~species) +
  labs(title = "Histogram of Penguin Bill Length by Species",
       x = "Flipper Length (mm)",
       y = "Frequency") +
  theme_minimal()
```



```
penguins %>%
  group_by(species) %>%
  summarise(mn = mean(flipper_length_mm, na.rm = TRUE),
            sd = sd(flipper_length_mm, na.rm = TRUE))
```

```
## # A tibble: 3 x 3
##   species      mn    sd
##   <fct>     <dbl> <dbl>
## 1 Adelie     190.  6.54
## 2 Chinstrap  196.  7.13
## 3 Gentoo     217.  6.48
```

```
penguins %>%
  filter(!is.na(sex)) %>%
  ggplot(aes(x = flipper_length_mm,
             y = bill_length_mm,
             col = species)) +
  geom_point() +
  geom_smooth(method = "lm",
              se = FALSE) +
  facet_wrap(~sex) +
  labs(title = "Bill Length vs. Flipper Length by Species",
       x = "Flipper Length (mm)",
       y = "Bill Length (mm)",
       col = "Species")
```



```r
adelie <- penguins %>%
  filter(species == "Adelie")

addmargins(table(adelie$sex, adelie$island))
```

```
##
##          Biscoe Dream Torgersen Sum
##   female     22    27        24  73
##   male       22    28        23  73
##   Sum        44    55        47 146
```

```r
table(adelie$island) / nrow(adelie)
```

```
##
##    Biscoe     Dream Torgersen
## 0.2894737 0.3684211 0.3421053
```

```r
adelie %>%
  group_by(island) %>%
  count()
```

```
## # A tibble: 3 x 2
## # Groups:   island [3]
##   island        n
##   <fct>     <int>
## 1 Biscoe       44
## 2 Dream        56
## 3 Torgersen    52
```

```
n <- nrow(adelie)
phat <- 44/nrow(adelie)

se_phat <- sqrt( phat * (1-phat) / n)

phat + c(-1, 1) * qnorm(0.975) * se_phat
```

```
## [1] 0.2173761 0.3615713
```

We are 95% confident that the true proportion of Adelie penguins who reside on Biscoe island is between 21.73% and 36.16%.

Suppose I have a random sample of 50 Adelie penguins. What is the probability that at most 25% of them are from Biscoe island?

Check CLT conditions:

1) Independence: penguins were randomly sampled

2) Large counts (success/failure): Expected success and failures are both greater than 10

```
n <- 50
n * phat
```

```
## [1] 14.47368
```

```
n*(1-phat)
```

```
## [1] 35.52632
```

```
mu_phat <- phat
se_phat <- sqrt(phat*(1-phat)/n)

pnorm(0.25, mean = mu_phat, sd = se_phat)
```

```
## [1] 0.2691263
```

Sampling Methods

- SRS: each sample of size n has the same probability of being chosen

- Stratified sampling: first group population based on characteristic, then take an SRS from each group

- Cluster sampling: group population into heterogenous clusters, and randomly select some entire clusters
- Systematic sampling: order population, then take every kth sample

- Voluntary response sampling: each person in the sample chooses to take part in the study

- Convenience sampling: participants are included in the study because they were easy to sample