Exploratory Data Analysis with the Tidyverse

一个关于企鹅的数据故事

诗与远方

2020-08-03

1 数据故事

今天讲一个关于企鹅的数据故事。数据来源这里，图片来源这里.

2 数据

2.1 导入数据

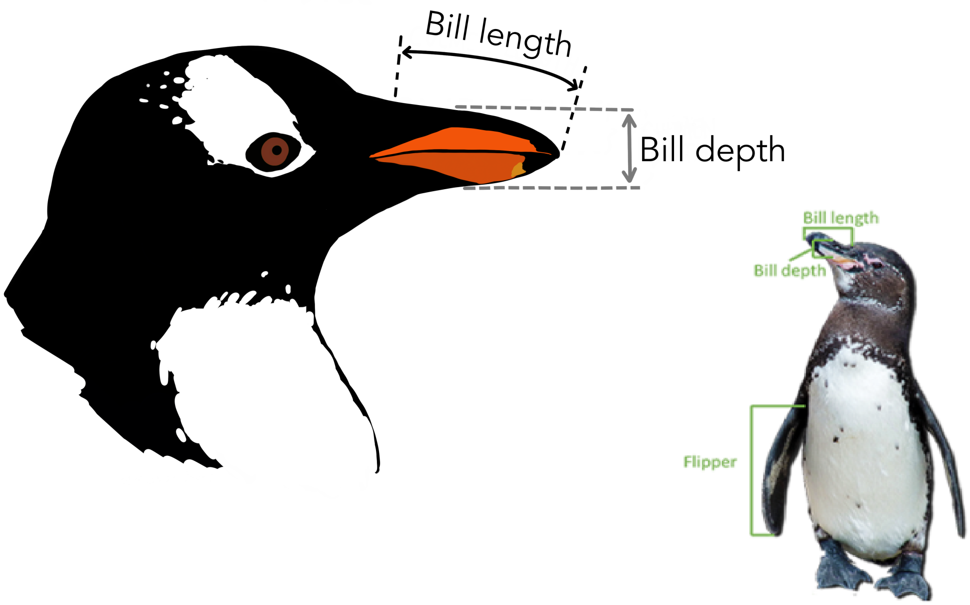
可通过宏包 palmerpenguins::penguins 获取数据，也可以读取本地 penguins.csv 文件，我们采取后面一种方法：

```
library(tidyverse)
penguins <- read_csv("./demo_data/penguins.csv")
penguins %>% head(5)
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	NA	NA	NA	NA	NA	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007

2.2 变量含义

variable	class	description
species	integer	企鹅种类 (Adelie, Gentoo, Chinstrap)
island	integer	所在岛屿 (Biscoe, Dream, Torgersen)
bill_length_mm	double	嘴峰长度 (单位毫米)
bill_depth_mm	double	嘴峰深度 (单位毫米)
flipper_length_mm	integer	鳍肢长度 (单位毫米)
body_mass_g	integer	体重 (单位克)
sex	integer	性别
year	integer	记录年份



2.3 数据清洗

```r
penguins %>% filter_all(any_vars(is.na(.)))
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	NA	NA	NA	NA	NA	2007
Adelie	Torgersen	34.1	18.1	193	3475	NA	2007
Adelie	Torgersen	42.0	20.2	190	4250	NA	2007
Adelie	Torgersen	37.8	17.1	186	3300	NA	2007
Adelie	Torgersen	37.8	17.3	180	3700	NA	2007
Adelie	Dream	37.5	18.9	179	2975	NA	2007
Gentoo	Biscoe	44.5	14.3	216	4100	NA	2007
Gentoo	Biscoe	46.2	14.4	214	4650	NA	2008
Gentoo	Biscoe	47.3	13.8	216	4725	NA	2009
Gentoo	Biscoe	44.5	15.7	217	4875	NA	2009
Gentoo	Biscoe	NA	NA	NA	NA	NA	2009

```r
d <- penguins %>% drop_na()
d %>% head()
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007
Adelie	Torgersen	38.9	17.8	181	3625	female	2007

3 探索性分析

3.1 多少种类的企鹅

```
d %>% count(species, sort = T)
```

species	n
Adelie	146
Gentoo	119
Chinstrap	68

3.2 多少个岛屿

```
d %>% count(island, sort = T)
```

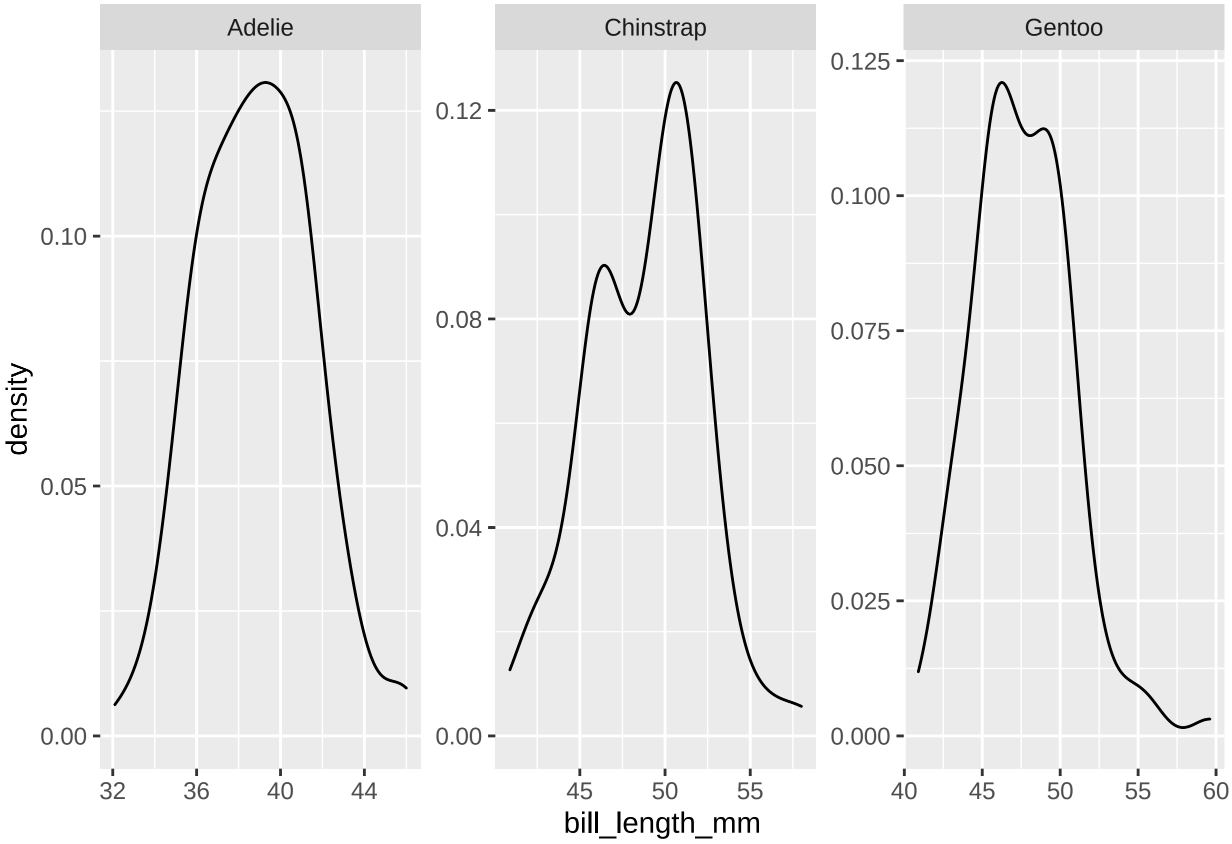
island	n
Biscoe	163
Dream	123
Torgersen	47

3.3 每种类型的企鹅，他们的各个属性的均值和分布

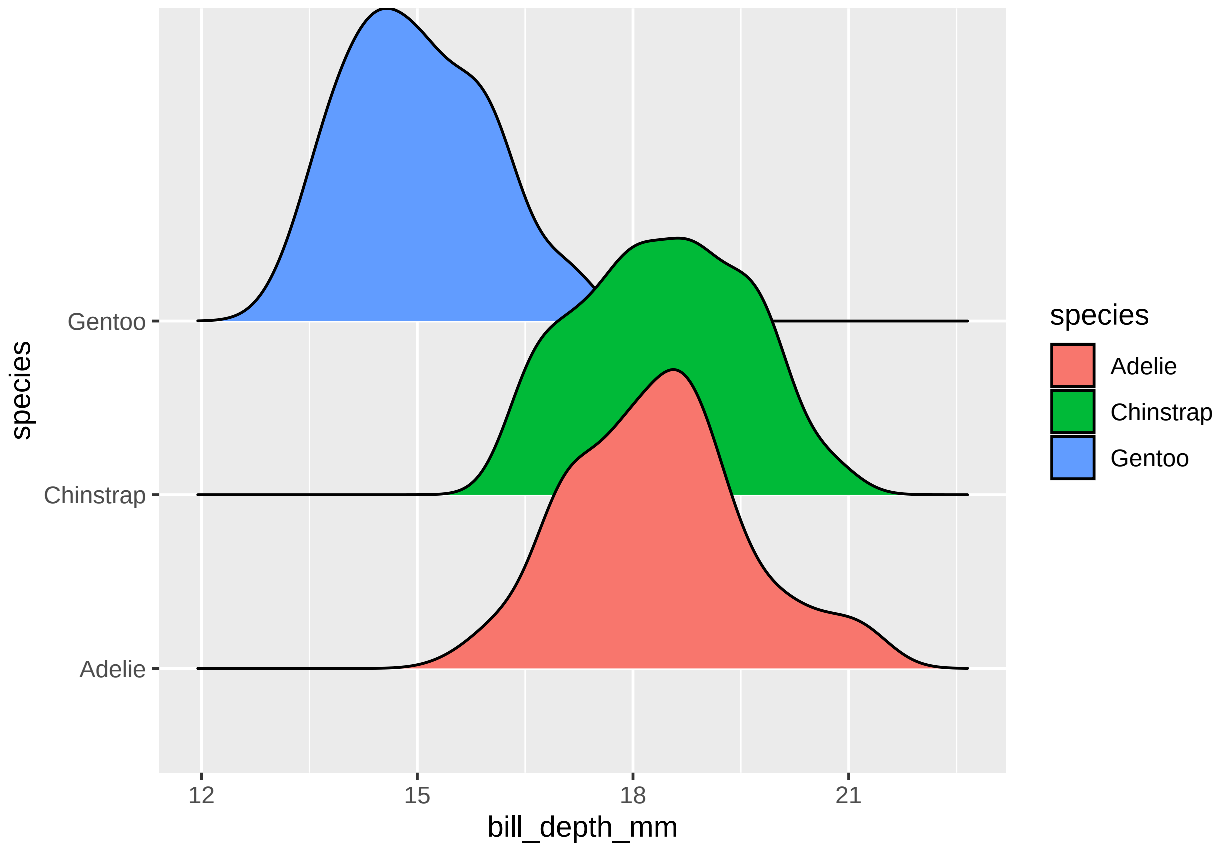
```
d %>%
  group_by(species) %>%
  summarise(
    across(where(is.numeric), mean, na.rm = T)
)
```

species	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	year
Adelie	38.82397	18.34726	190.1027	3706.164	2008.055
Chinstrap	48.83382	18.42059	195.8235	3733.088	2007.971
Gentoo	47.56807	14.99664	217.2353	5092.437	2008.067

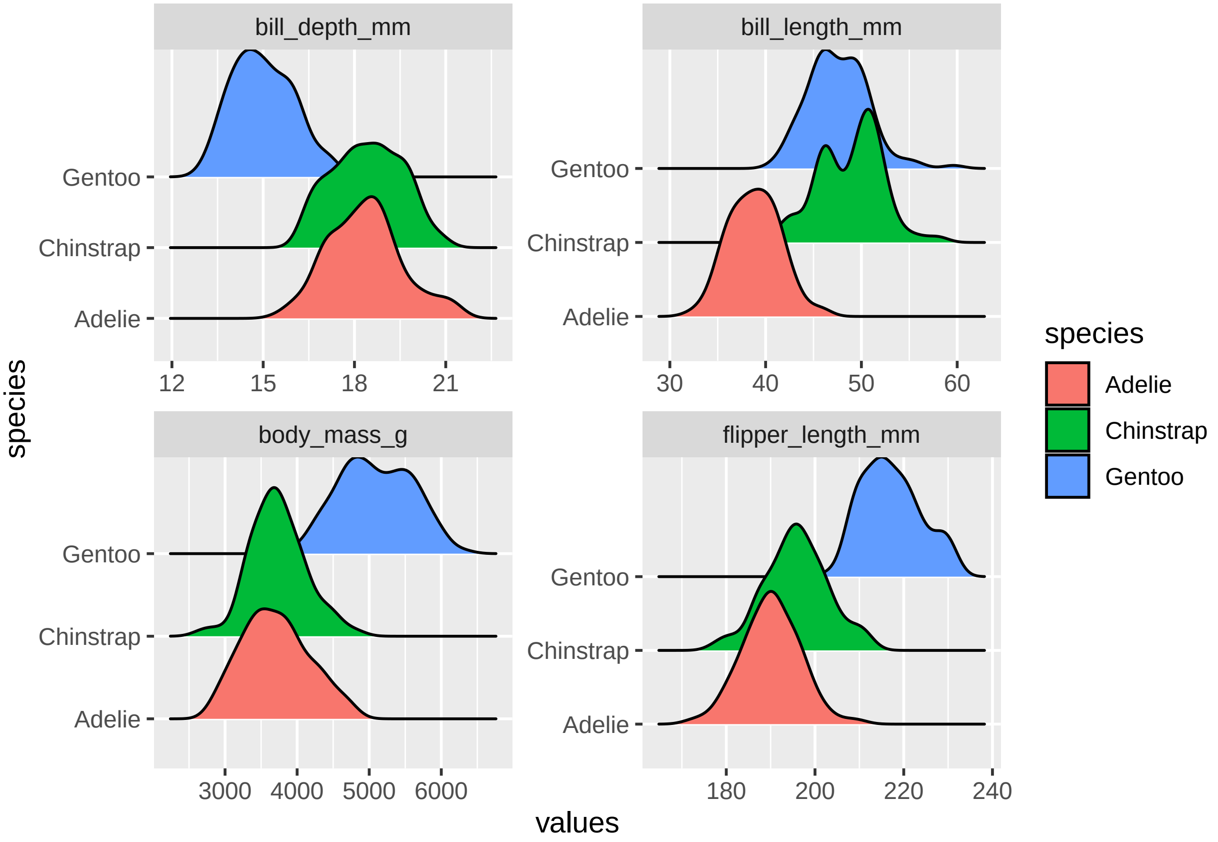
```
d %>%
  ggplot(aes( x = bill_length_mm)) +
  geom_density() +
  facet_wrap(vars(species), scale = "free")
```



```
library(ggridges)
d %>%
  ggplot(aes( x = bill_depth_mm, y = species, fill = species) ) +
  ggridges::geom_density_ridges()
```

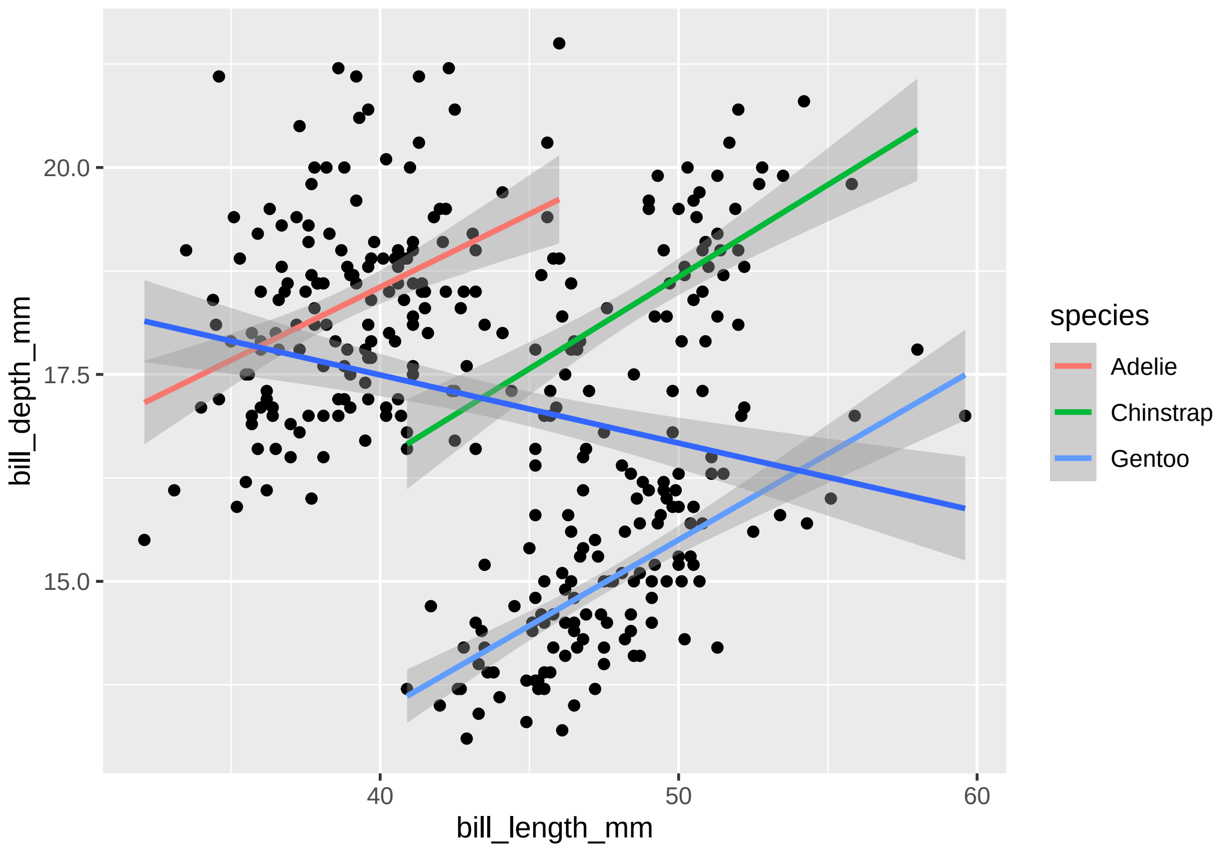


```r
d %>% select(species, body_mass_g, ends_with("_mm")) %>%
  pivot_longer(
    cols = -species,
    names_to = "metric",
    values_to = "values"
  ) %>%
  ggplot(aes(x = values, y = species, fill = species) ) +
  ggridges::geom_density_ridges() +
  facet_wrap(vars(metric), scale = "free")
```



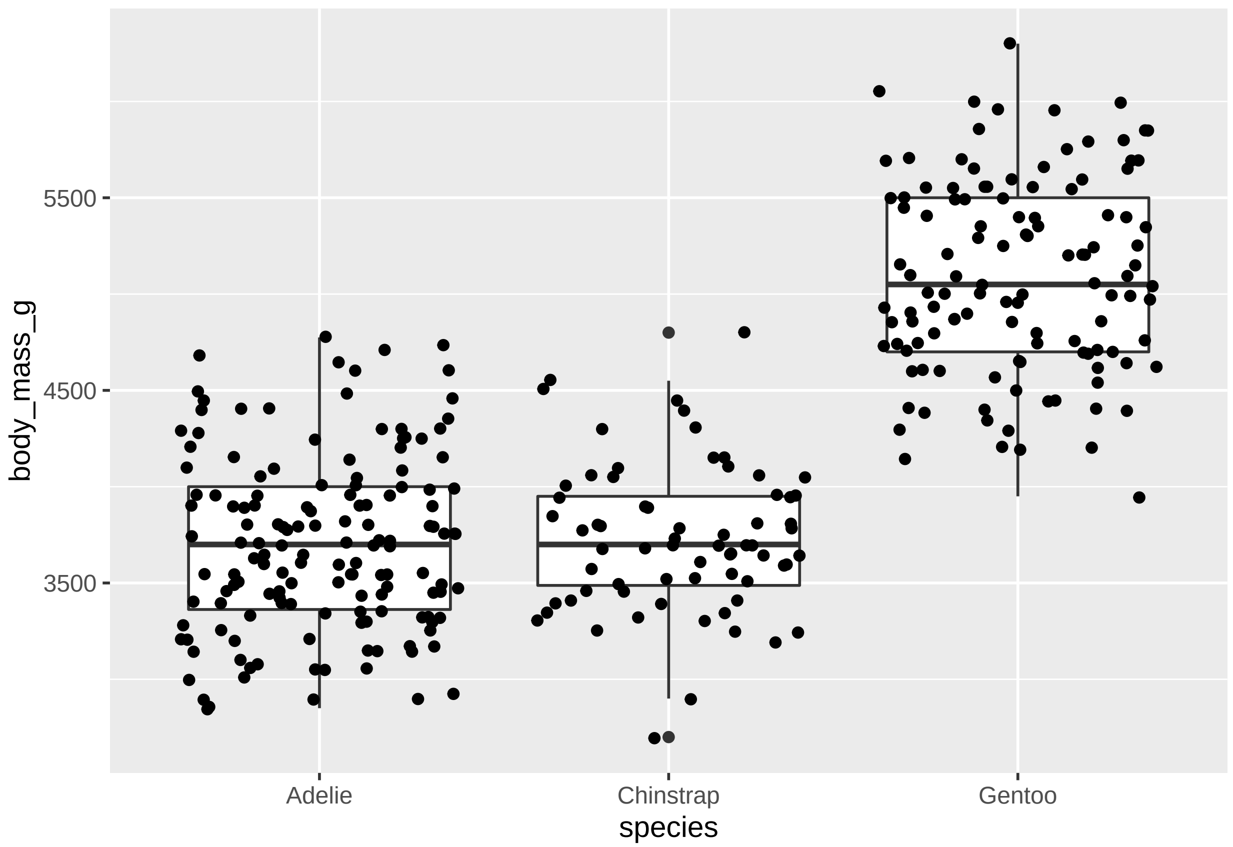
3.4 嘴巴的长度和深度的关联？

```
d %>%
  ggplot(aes(x = bill_length_mm, y = bill_depth_mm)) +
  geom_point() +
  geom_smooth(method = lm, aes(color = species)) +
  geom_smooth(method = lm)
```



3.5 不同种类的宝宝，体重具有显著性差异？

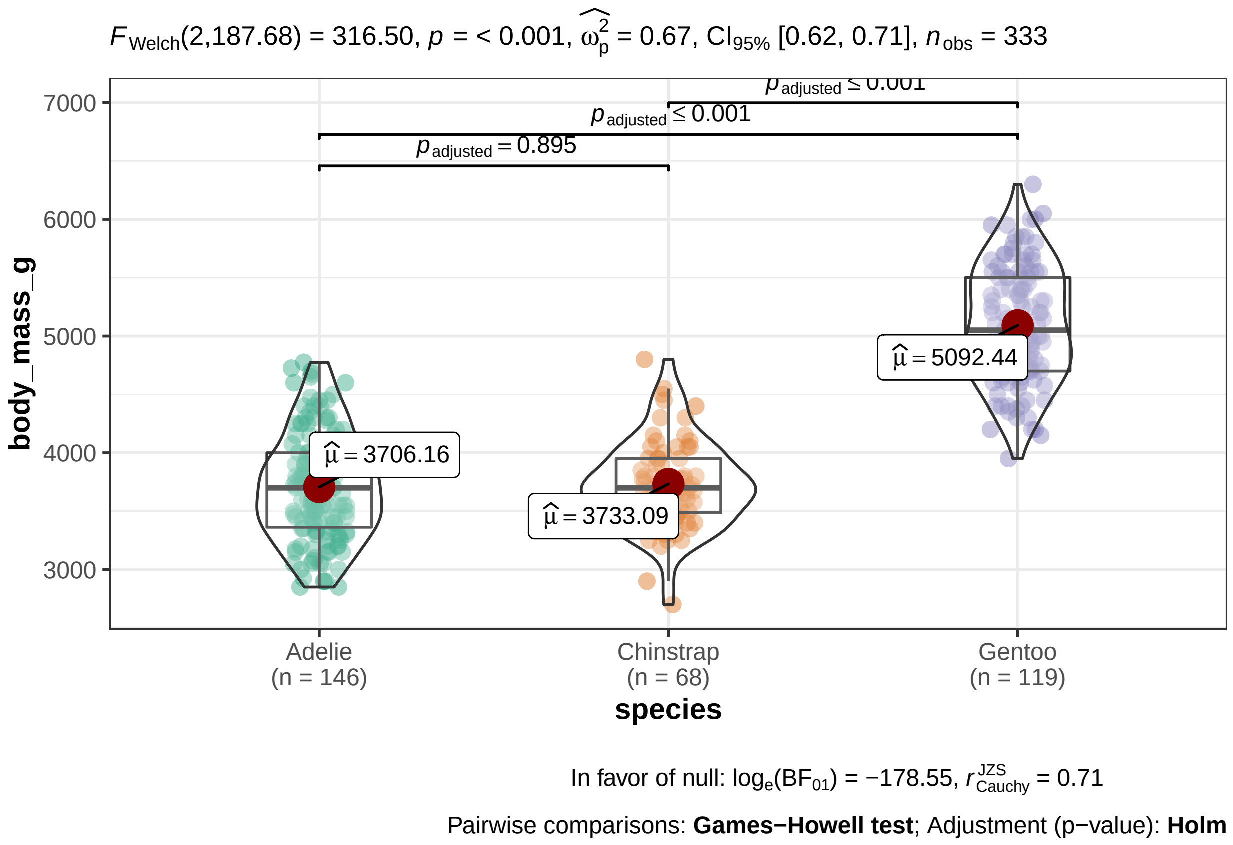
```
d %>%
  ggplot(aes(x = species, y = body_mass_g)) +
  geom_boxplot() +
  geom_jitter()
```



```
aov(body_mass_g ~ species, data = d) %>% summary()
```

```
##              Df    Sum Sq  Mean Sq F value Pr(>F)
## species       2 145190219 72595110   341.9 <2e-16 ***
## Residuals   330  70069447   212332
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
library(ggstatsplot)
d %>%
  ggbetweenstats(
    x = species,
    y = body_mass_g,
    pairwise.comparisons = T,
    pairwise.display = T
  )
```

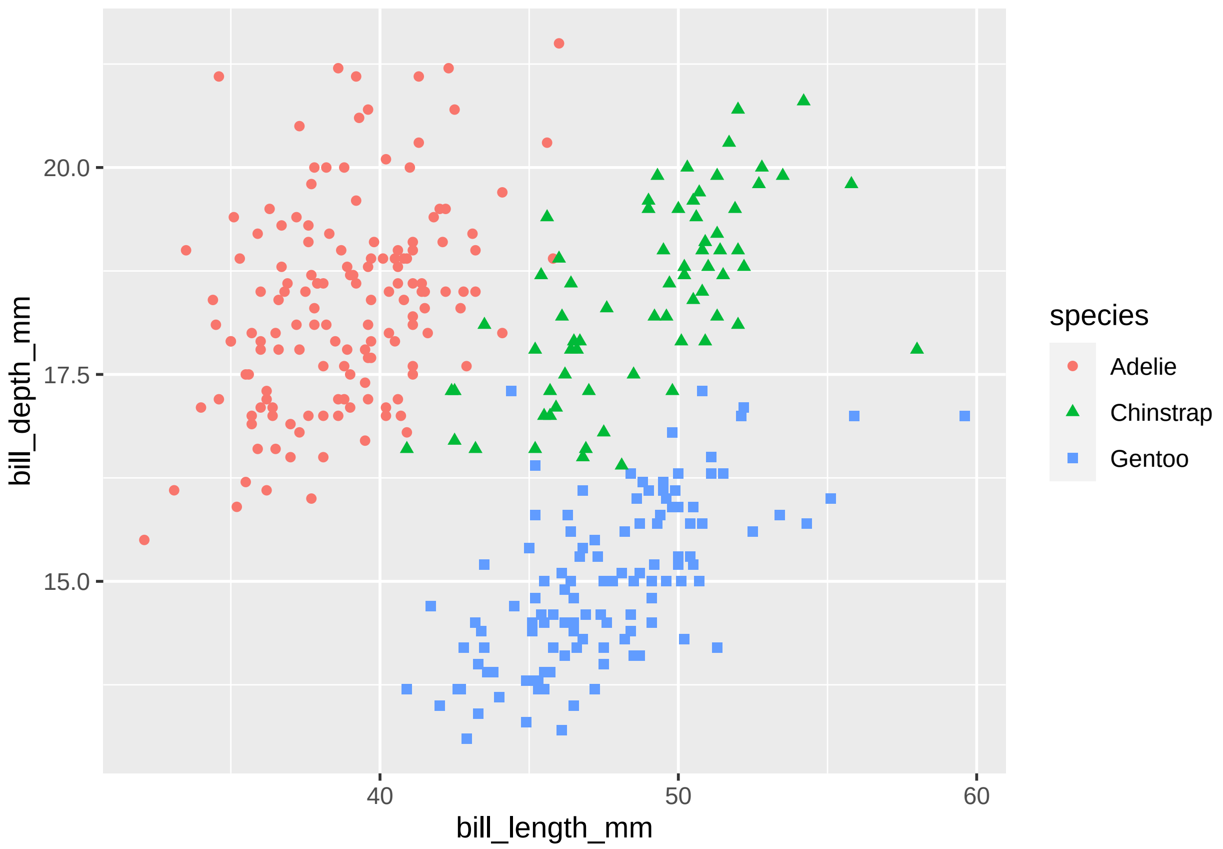


使用这个宏包辅助我们学习统计

3.6 通过嘴巴的长度和深度，区分企鹅的种类？性别？

这是机器学习的范畴

```
d %>%
  ggplot(aes(x = bill_length_mm, y = bill_depth_mm, color = species, shape = species)) +
  geom_point()
```



```
library(tidymodels)
d <- d %>% mutate(species = factor(species))

    split <- initial_split(d)
    split
```

```
## <Analysis/Assess/Total>
## <250/83/333>
```

```
   training_data <-  training(split)
  testing_data <-   testing(split)

 model <- parsnip::nearest_neighbor() %>%
    set_engine("kknn") %>%
    set_mode("classification") %>%
    fit(species ~ bill_length_mm + bill_depth_mm, data = training_data)


 predict(model, new_data = testing_data) %>%
```

```r
  bind_cols(testing_data) %>%
  count(species, .pred_class)
```

species	.pred_class	n
Adelie	Adelie	40
Adelie	Chinstrap	2
Chinstrap	Chinstrap	14
Gentoo	Chinstrap	1
Gentoo	Gentoo	26