FEIT EDUCATION 飞天教育

— 选择飞天，重仓未来 —

# Week 9 知识梳理 – DATA2002

by

Kathleen老师

FEIT EDUCATION
飞天教育

FEIT小助手【USYD】
澳大利亚 悉尼

## 今日课程内容分布

Github 使用
DATA2002 Project Group Contract & EDA

## GitHub 演示

现场演示

https://stackoverflow.com/questions/32699891/rstudio-push-rpostback-askpass-error

## Group Contract 1% & EDA 1%

### Group Contract 1%

Due: 9 Oct 2022, W9

要求:

- 标红的三点直接复制粘贴,必须保留
- 整个小组再想至少四点规则,大家一起遵守,每个人必须一样才能拿分

所有组员必须单独提交自己那份 contract，未提交的组员自动判定为 0 分

例子:

#### Group contract

**Group number: CC901E9**
**Name: Garth Tarr**
**GitHub link: https://github.sydney.edu.au/gtar4178/CC901E9**
**I agree to:**

- **Abide by the terms of this contract in relation to the group assessment for DATA2002/2902.**
- **Store all my contributions to the assessment in the GitHub repository.**
- **Keep an accurate record of my contribution to the assessment. A copy of this may be requested by the coordinator.**
- *Work cooperatively, treat each other with respect, act honestly and ethically and not engage in any activities that could be perceived as bullying or harassment, as detailed in the Student Contract*

- *Communicate in two main ways: informal discussions on Slack and using the "Issues" functionality on GitHub to provide updates on specific tasks, including tagging responsibility to specific group members.*
- *Check Slack daily and check in with GitHub at least once a week and more regularly as we get closer to the deadline. If something on GitHub is urgent, it will be highlighted in Slack.*
- *Attend labs in the weeks before the tasks are due and meet for lunch on the day of the lab to give us time to informally discuss any issues we're facing. Other meetings will be held via Zoom and arranged on an ad hoc basis.*

I understand that:

- My agreement to these terms is indicated through the act of submitting this in Canvas.
- If I fail to meet my obligations as detailed in this group contract, then I have failed to meet the assessment requirements for DATA2002/2902 and may be awarded a mark of zero for some or all of the project components.

## EDA 1%

要求:

- 读取 dataset
- 选择变量
- 生成一些相对应的 graph

例子:

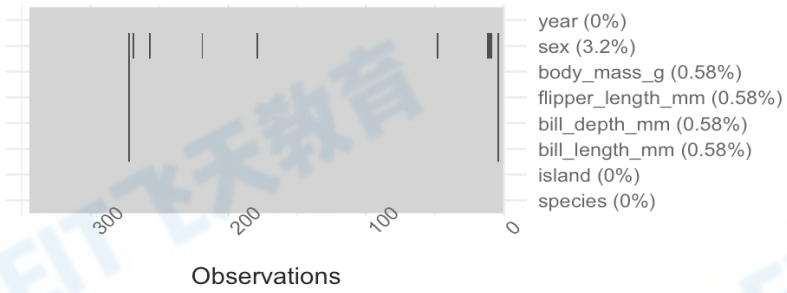
## Exploratory data analysis

**Data set:** Palmer penguins
**Dependent variable:** Body mass

```
library(tidyverse)
theme_set(theme_bw())
data("penguins", package = "palmerpenguins")
```

Looking at the pattern of missing data:

```
visdat::vis_miss(penguins) + coord_flip() + theme(legend.position = "none
")
```


year (0%)
sex (3.2%)
body_mass_g (0.58%)
flipper_length_mm (0.58%)
bill_depth_mm (0.58%)
bill_length_mm (0.58%)
island (0%)
species (0%)
300
200
100
0
Observations

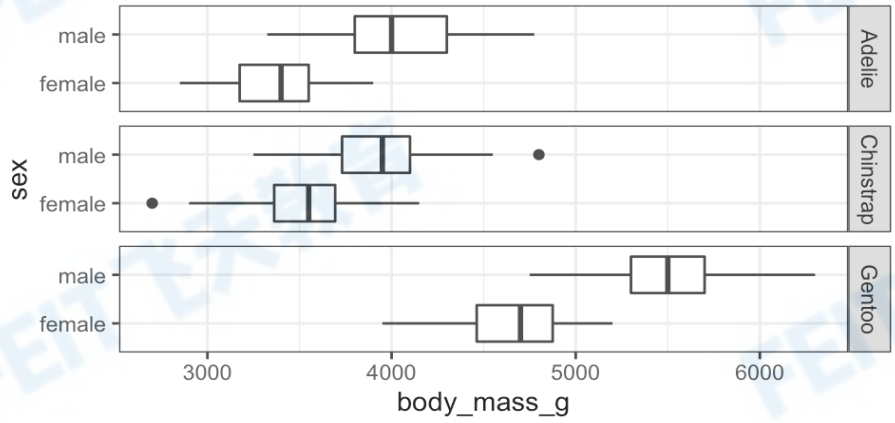
Descriptive statistics:

```
table1::table1(~ bill_length_mm + bill_depth_mm + flipper_length_mm + body
_mass_g + island | species + sex, data = penguins, overall = FALSE)
```

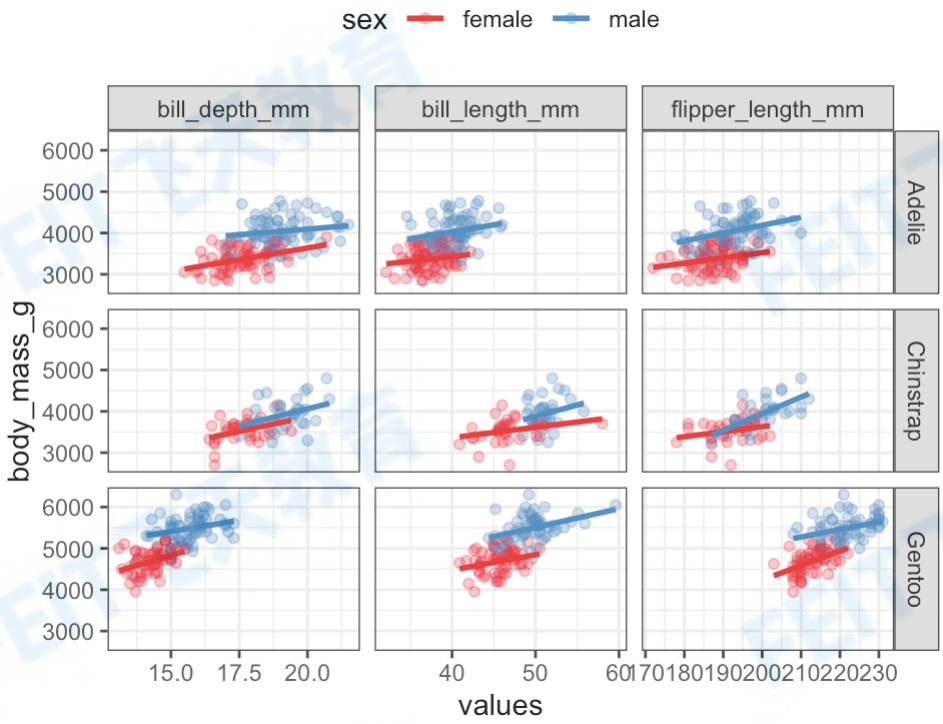
| | Adelie | | Chinstrap | | Gentoo | |
|---|---|---|---|---|---|---|
| | female (N=73) | male (N=73) | female (N=34) | male (N=34) | female (N=58) | male (N=61) |
| **bill_length_mm** | | | | | | |
| Mean (SD) | 37.3 (2.03) | 40.4 (2.28) | 46.6 (3.11) | 51.1 (1.56) | 45.6 (2.05) | 49.5 (2.72) |
| Median [Min, Max] | 37.0 [32.1, 42.2] | 40.6 [34.6, 46.0] | 46.3 [40.9, 58.0] | 51.0 [48.5, 55.8] | 45.5 [40.9, 50.5] | 49.5 [44.4, 59.6] |
| **bill_depth_mm** | | | | | | |
| Mean (SD) | 17.6 (0.943) | 19.1 (1.02) | 17.6 (0.781) | 19.3 (0.761) | 14.2 (0.540) | 15.7 (0.741) |
| Median [Min, Max] | 17.6 [15.5, 20.7] | 18.9 [17.0, 21.5] | 17.7 [16.4, 19.4] | 19.3 [17.5, 20.8] | 14.3 [13.1, 15.5] | 15.7 [14.1, 17.3] |
| **flipper_length_mm** | | | | | | |
| Mean (SD) | 188 (5.60) | 192 (6.60) | 192 (5.75) | 200 (5.98) | 213 (3.90) | 222 (5.67) |
| Median [Min, Max] | 188 [172, 202] | 193 [178, 210] | 192 [178, 202] | 201 [187, 212] | 212 [203, 222] | 221 [208, 231] |
| **body_mass_g** | | | | | | |
| Mean (SD) | 3370 (269) | 4040 (347) | 3530 (285) | 3940 (362) | 4680 (282) | 5480 (313) |
| Median [Min, Max] | 3400 [2850, 3900] | 4000 [3330, 4780] | 3550 [2700, 4150] | 3950 [3250, 4800] | 4700 [3950, 5200] | 5500 [4750, 6300] |
| **island** | | | | | | |
| Biscoe | 22 (30.1%) | 22 (30.1%) | 0 (0%) | 0 (0%) | 58 (100%) | 61 (100%) |
| Dream | 27 (37.0%) | 28 (38.4%) | 34 (100%) | 34 (100%) | 0 (0%) | 0 (0%) |
| Torgersen | 24 (32.9%) | 23 (31.5%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |

Initial visualisations:

```
penguins |> drop_na() |> ggplot() + aes(y = sex, x = body_mass_g) +
  geom_boxplot() + facet_grid( species ~ .)
```




```
penguins |> select(-island, - year) |>
  pivot_longer(cols = c(bill_length_mm, bill_depth_mm, flipper_length_mm),
               names_to = "variable", values_to = "values") |>
  drop_na() |> ggplot() +
  aes(x = values, y = body_mass_g, colour = sex) +
  geom_point(alpha = 0.3) +
  facet_grid(species ~ variable, scales = "free_x") +
  scale_colour_brewer(palette = "Set1") +
  geom_smooth(method = "lm", se = FALSE) +
  theme(legend.position = "top")
```




Summary: From my initial EDA it looks like species and sex are both important factors for predicting body mass. There also appears to be relationships between body mass and the bill and flipper measurements.

---

*Note: your EDA doesn't necessarily need to include the R code, I included mine just to show how I generated the outputs. I still used a .qmd file but then copied the output into this Word document. If you're interested, the start of the .qmd file was:*

```
---
title: 'Project EDA'
format:
  docx:
    fig-format: retina
---

```{r, message=FALSE}
library(tidyverse)
theme_set(theme_bw())
data("penguins", package = "palmerpenguins")
```
```

# Lecture 25 Two-factor ANOVA with interactions

## Main effects: interpretation

- So each $\alpha_i$ in fact is a contrast in the (treatment combination) group means that measures in some overall sense how the means for level $i$ of factor A differ from the overall average.
- In exactly the same way, each $\gamma_j$ is a contrast (in the $\mu_{ij}$'s) that compares (in some overall sense) level $j$ of factor B to the overall average.

## Interaction effects

- A similar (but more complicated) calculation can be used to show that each $(\alpha\gamma)_{ij}$ is also a contrast in the $\mu_{ij}$'s.
- Each

$$(\alpha\gamma)_{ij} = \mu_{ij} - (\mu + \alpha_i + \gamma_j)$$

  compares a mean $\mu_{ij}$ to the corresponding "additive prediction" $\mu + \alpha_i + \gamma_j$.
  - If the factor levels *actually do* combine additively, then each such interaction (population) contrast is zero.
- Therefore a "test for no interaction" can be formulated as the following null hypothesis:

$$H_0\text{: } (\alpha\gamma)_{ij} = 0 \text{ for all } i, j.$$

- The interaction terms are such that for each fixed $i$,

$$\sum_{j=1}^{b} (\alpha\gamma)_{ij} = 0$$

  and vice versa for each fixed $j$.
- There are thus $(a-1)(b-1)$ "free" interaction effects (like in an $a$ by $b$ two-way contingency table where all row and column sums are fixed).
- That is to say, there are $(a-1)(b-1)$ degrees of freedom for the interaction effects.

例 4:

研究 3 种毒药与 4 种解药之间对生存时间的关系

```
# library(BHH2)
data("poison.data", package = "BHH2")

poison.data = poison.data %>% rename(antidote = treat)

poison.data = poison.data %>% mutate(inv_survival = 1/y)

poison.data %>% ggplot() +
aes(x = poison, y = inv_survival) + geom_boxplot() +
theme_classic(base_size = 30) + facet_wrap(~antidote, ncol =
4) + labs(y = "1/Survival time")

summary(aov(inv_survival ~ poison * antidote, data =
poison.data))
```

例 5:
研究纸质与折法对纸飞机飞行距离的影响

```
planes = read_tsv("https://raw.githubusercontent.com/DATA2002/data/master/planes.txt")

planes = planes %>% mutate(
Paper = case_when( Paper == 1 ~ "80gsm", Paper == 2 ~ "50gsm"
),
Plane = case_when(
Plane == 1 ~ "HiPerf",
Plane == 2 ~ "Simple" )
)

p1 = ggplot(planes, aes(x = Plane, y = Distance)) +
geom_boxplot() + labs(y = "Distance (mm)")
p2 = ggplot(planes, aes(x = Paper, y = Distance)) +
geom_boxplot() + labs(y = "Distance (mm)")
gridExtra::grid.arrange(p1, p2, ncol = 2)

plane_aov = aov(Distance ~ Paper * Plane, data = planes)
summary(plane_aov)
```

## Interaction plots

```
emmip(plane_aov, Plane ~ Paper) + theme_classic(base_size = 36)

emmip(plane_aov, Paper ~ Plane) + theme_classic(base_size = 36)
```

## Post hoc comparisons no significant interaction

```
a2 = aov(inv_survival ~ poison + antidote, data = poison.data)

p_emm = contrast(emmeans(a2, ~poison), method = "pairwise", adjust = "bonferroni")

p_emm

a_emm = contrast(emmeans(a2, ~antidote), method = "pairwise", adjust = "bonferroni")

a_emm

pa_emm = update(p_emm + a_emm)
```

### Post hoc comparisons with significant interaction

```
plane_aov = aov(Distance ~ Plane*Paper, data = planes)

plane_emm = emmeans(plane_aov, ~ Paper + Plane)

contrast(plane_emm, method = "pairwise", adjust = "tukey")
```

## Lecture 26 Simple linear regression

### Supervised learning vs. Unsupervised learning

近些年大火的 Machine learning 中，所谓 machine 所在做的 learning 可以粗略的分为两
类- Supervised learning 和 Unsupervised learning。

**Supervised learning**

- We have knowledge of class labels or values.
- Goal: train a model using known class labels to predict class or value label for a new data point.

**Unsupervised learning**

- No knowledge of output class or value – data is unlabelled.
- Goal: determine data patterns/groupings.



Supervised learning 是基于已知的 class label 来进行学习，而 unsupervised learning 是根据未知的 class label 来进行学习。

Supervised learning 的目的更多的是来【根据给定数据预测新数据】，也就是说我们基
于我们所给定的数据，搭配上一个我们自己演算出来的【model】，来预测【新数据在
模型中对应的值】。比如说，我们根据房子的地段，大小，本地人均收入等来推断一个新上市的房子的成交价格。

反之，Unsupervised learning 是根据【未知特性的数据对其进行分类】。
比如说，我们将 10 个人，每个人 10 张照片，共 100 张照片的数据分别打散，然后试图根据 unsupervised learning 来归类出【这 10 张是谁的照片】。

基于以上所说，我们这次 Report 将要进行的将是基于【Supervised learning】。

而Supervised learning 中又分为两个分支：Regression 和 Classification

- Classification 是对数据进行【Classify】行为，故他的output 将会是一个class label
- Regression 的output 则是【Generate continuous output】，也就是类似在预测房价时输出一个【具体的数字】

## Regression

我们已经知道了，regression 是根据数据来预测数据，所以我们能够联想到我们最常用
的【根据数据来预测数据】就是在y = ax+b 这类的方程中替换进一些对应的数值。

而我们所谓的Simple linear regression model，本质上就是在建立一个y=ax+b 的方程，然后在过程中我们通过调试a 以及b 的值来使得这条线能够最大的贴合我们的数据。

A **simple linear regression** model aims to predict an outcome variable, $Y$, using a single predictor variable $x$,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

for $i = 1, 2, \ldots, n$ where $n$ is the number of observations (rows) in the data set.

This is just the equation of a straight line (like $y = mx + b$) plus some additional variation,

- $\beta_0$ is the population intercept parameter
- $\beta_1$ is the population slope parameter
- $\varepsilon_i$ is the error term and typically assumed to follow $N(0, \sigma^2)$

## How to estimate?

We aim to minimise the sum of squared residuals.

- What's a residual?

$$r_i = y_i - \hat{y}_i$$

where $\hat{y}_i$ is the fitted value, the value we predict for the $i$th observation given the $i$th predictor value:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

在 R 当中，我们可以通过

```
lm(y ~ x, data)
```

来一行得到我们的 model

我们如果要建立一个Simple linear regression model 的话， 我们需要进行四个 assumption：

- Linearity - the relationship between and is linear
- Independence - all the errors are independent of each other
- Homoskedasticity - the errors have constant variance for all
- Normality - the errors follow a normal distribution

```
library(caret)
cv_full = train(
  lozone ~ radiation + temperature + wind, environmental,
  method = "lm",
  trControl = trainControl(
    method = "cv", number = 10,
    verboseIter = FALSE
  )
)
cv_full
```

```
## Linear Regression
##
## 111 samples
##   3 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 100, 100, 99, 100, 101, 99, ...
## Resampling results:
##
##   RMSE       Rsquared  MAE
##   0.5110818  0.697293  0.3990078
```

```
cv_simple = train(
  lozone ~ temperature,
  environmental,
  method = "lm",
  trControl = trainControl(
    method = "cv", number = 10,
    verboseIter = FALSE
  )
)
cv_simple
```

```
## Linear Regression
##
## 111 samples
##   1 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 99, 99, 100,
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.5639179  0.5597842  0.4403246
```