Modern Data Science Methods for Educational Research

R for Data Analysis in Educational Research

Data Analysis

ผศ.ดร.สิวะโชติ ศรีสุทธิยากร
อ.ดร.ประภาศิริ รัชประภาพรกุล

ภาควิชาวิจัยและจิตวิทยาการศึกษา
คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

February 12, 2023


KRUROO
EDU @ CHULALONGKORN

1. Exploratory Data Analysis

Exploratory Data Analysis (EDA)

EDA เป็นกระบวนการวิเคราะห์เพื่อสำรวจข้อมูล ที่มีลักษณะเป็นกระบวนการทวนซ้ำ โดยมีวัตถุประสงค์เพื่อ

- ตั้งคำถาม/ประเด็นวิจัยจากข้อมูล
- **หาคำตอบที่เหมาะสมของคำถาม**ด้วยการใช้การจัดเตรียมข้อมูล ร่วมกับการทำทัศนภาพข้อมูล และ/หรือ การใช้สถิติวิเคราะห์
- ผลการวิเคราะห์ที่ได้อาจนำไปสู่ประเด็นคำถามวิจัยใหม่ ๆ

Exploratory Data Analysis (EDA) ในกระบวนการวิเคราะห์ข้อมูล

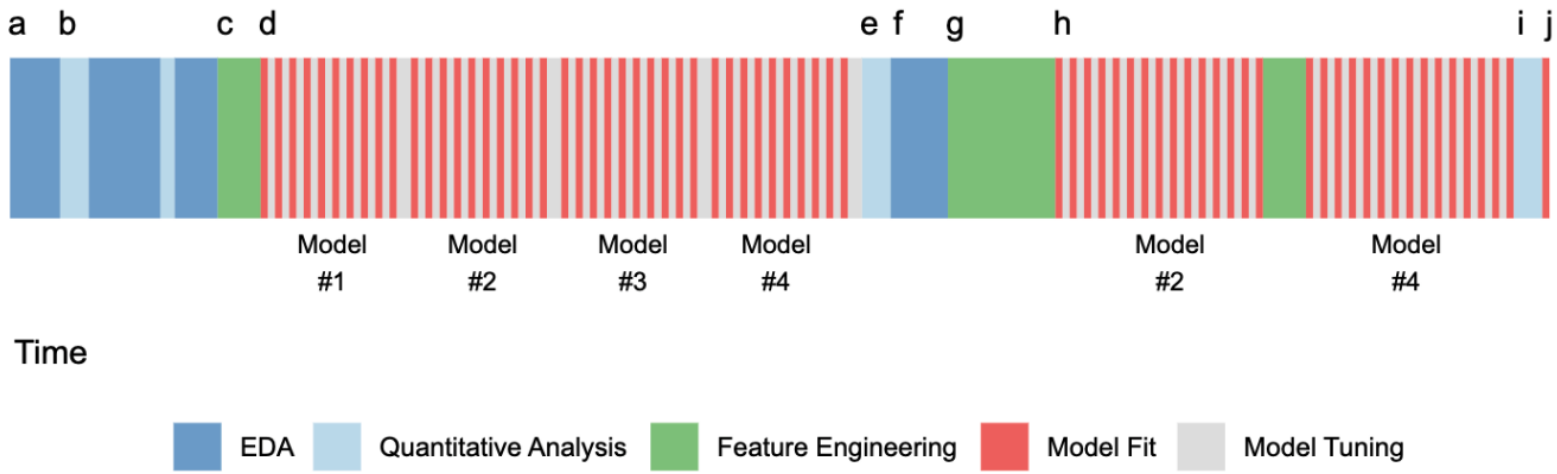


Figure 1: ที่มา : Max Khun, & Kjell Johnson (2019)

Exploratory Data Analysis (EDA)

- การสำรวจโครงสร้างชุดข้อมูล - `str()`/`glimpse()`, `head()`, `tail()`
- การสำรวจสภาพของตัวแปร
- การสำรวจความสัมพันธ์ระหว่างตัวแปร

ชุดข้อมูลที่ใช้เป็นตัวอย่าง

เนื้อหาส่วนนี้จะใช้ชุดข้อมูล `TeacherSalaryData.csv` เป็นตัวอย่างประกอบเนื้อหา

```
1 library(dplyr)
2 dat <- read.csv("TeacherSalaryData.csv",
3                 header = TRUE,
4                 stringsAsFactors = TRUE)
```

วัตถุประสงค์ของการวิจัย

1. (เชิงเปรียบเทียบ) เพื่อเปรียบเทียบค่าเฉลี่ยเงินเดือนของอาจารย์มหาวิทยาลัย ระหว่าง กลุ่มอาจารย์ที่มี ตำแหน่งทางวิชาการ สาขาวิชา และเพศแตกต่างกัน
2. (วิเคราะห์ความสัมพันธ์)
   เพื่อวิเคราะห์ความสัมพันธ์ระหว่างเงินเดือนของอาจารย์มหาวิทยาลัย กับ ตำแหน่งทางวิชาการ สาขาวิชา เพศ และประสบการณ์ทำงานที่แตกต่างกัน

สำรวจโครงสร้างของชุดข้อมูล

```
1 glimpse(dat)
```

```
Rows: 397
Columns: 7
$ X              <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
$ rank           <fct> Prof, Prof, AsstProf, Prof, Prof, Ass
$ discipline     <fct> B, B, B, B, B, B, B, B, B, B, B, B, B
$ yrs.since.phd  <int> 19, 20, 4, 45, 40, 6, 30, 45, 21, 18,
$ yrs.service    <int> 18, 16, 3, 39, 41, 6, 23, 45, 20, 18,
$ sex            <fct> Male, Male, Male, Male, Male, Male, M
$ salary         <int> 139750, 173200, 79750, 115000, 141500
```

2. สำรวจการแจกแจงของตัวแปร

การสำรวจการแจกแจงของตัวแปร

- สำรวจขอบเขต ค่าที่เป็นไปได้ของตัวแปรแต่ละตัว ความสมเหตุสมผลของค่าดังกล่าว
- สำรวจการกระจายของตัวแปรแต่ละตัว
- ฟังก์ชันที่อาจนำมาใช้ในการดำเนินงาน `summary()`, `hist()`, `boxplot()`, `barplot()` , ...

สำรวจการแจกแจงของตัวแปร : สถิติพื้นฐาน

```
1 summary(dat)
```

```
        rank     discipline yrs.since.phd    yrs.service
 AssocProf: 64   A:181      Min.   : 1.00   Min.   : 0.00
 AsstProf : 67   B:216      1st Qu.:12.00   1st Qu.: 7.00
 Prof     :266              Median :21.00   Median :16.00
                            Mean   :22.31   Mean   :17.61
                            3rd Qu.:32.00   3rd Qu.:27.00
                            Max.   :56.00   Max.   :60.00

     salary
 Min.   : 57800
 1st Qu.: 91000
 Median :107300
 Mean   :113706
 3rd Qu.:134185
```

สำรวจการแจกแจงของตัวแปร : สถิติพื้นฐาน

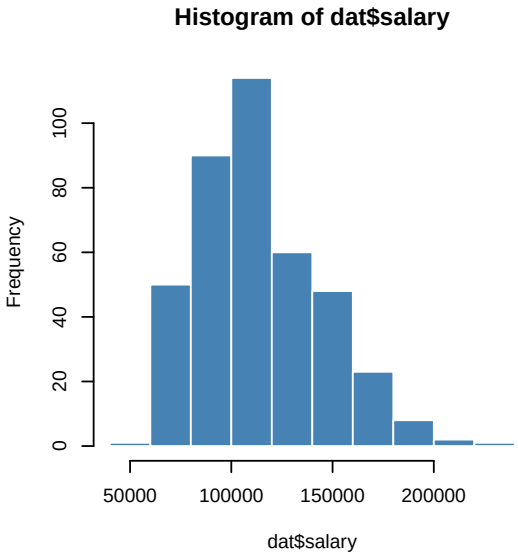
```
# install.packages("psych")
library(psych)
# calculate descriptive stat
describe(dat)
```

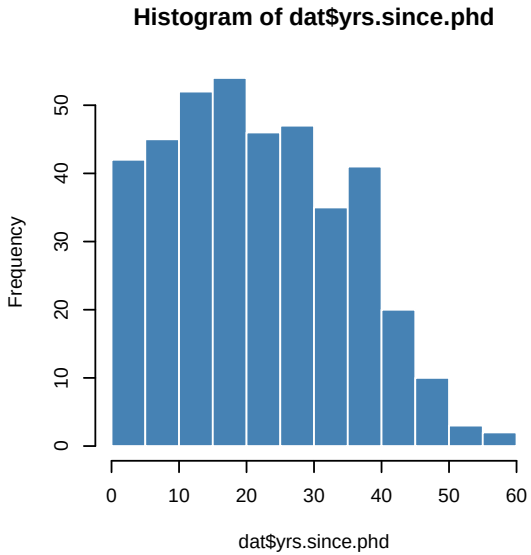
```
                vars   n       mean         sd median    trimmed
rank*              1 397       2.51       0.76      3       2.63
discipline*        2 397       1.54       0.50      2       1.55
yrs.since.phd      3 397      22.31      12.89     21      21.83
yrs.service        4 397      17.61      13.01     16      16.51
sex*               5 397       1.90       0.30      2       2.00
salary             6 397  113706.46  30289.04 107300  111401.61
                  max  range  skew kurtosis      se
rank*               3      2 -1.14    -0.31    0.04
discipline*         2      1 -0.18    -1.97    0.03
```

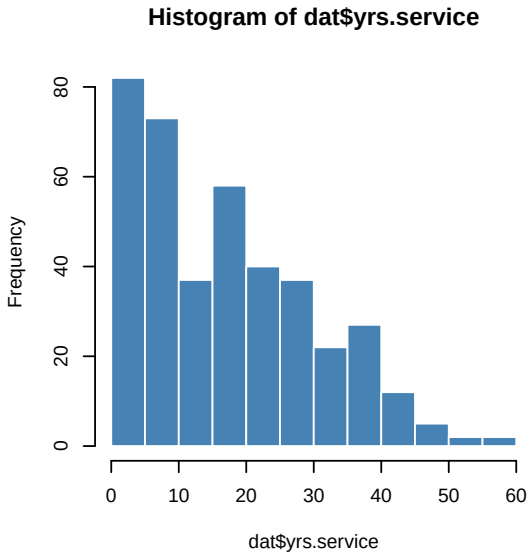
สำรวจการแจกแจงของตัวแปร : ตัวแปรต่อเนื่อง

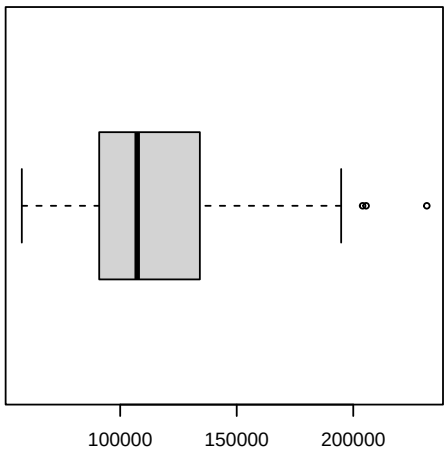
```
par(mfrow=c(2,3))
hist(dat$salary, col="steelblue", border="white")
hist(dat$yrs.since.phd, col="steelblue", border="white")
hist(dat$yrs.service, col="steelblue", border="white")
boxplot(dat$salary, horizontal = TRUE)
boxplot(dat$yrs.since.phd, horizontal = TRUE)
boxplot(dat$yrs.service, horizontal = TRUE)
```

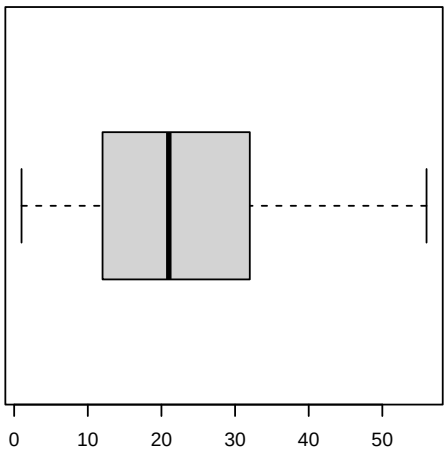
สำรวจการแจกแจงของตัวแปร : ตัวแปรต่อเนื่อง

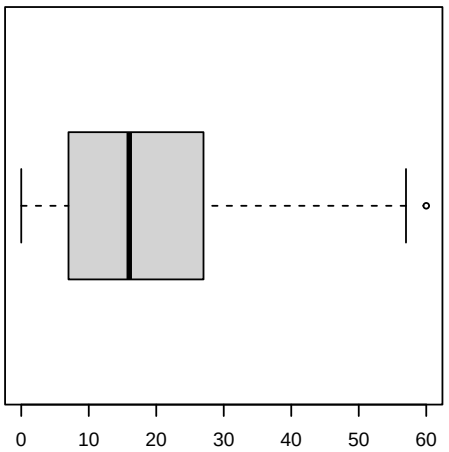












สำรวจการแจกแจงของตัวแปร : ตัวแปรจัดประเภท

```
tab_rank <- table(dat$rank)
tab_rank
```

```
AssocProf  AsstProf      Prof
       64        67       266
```

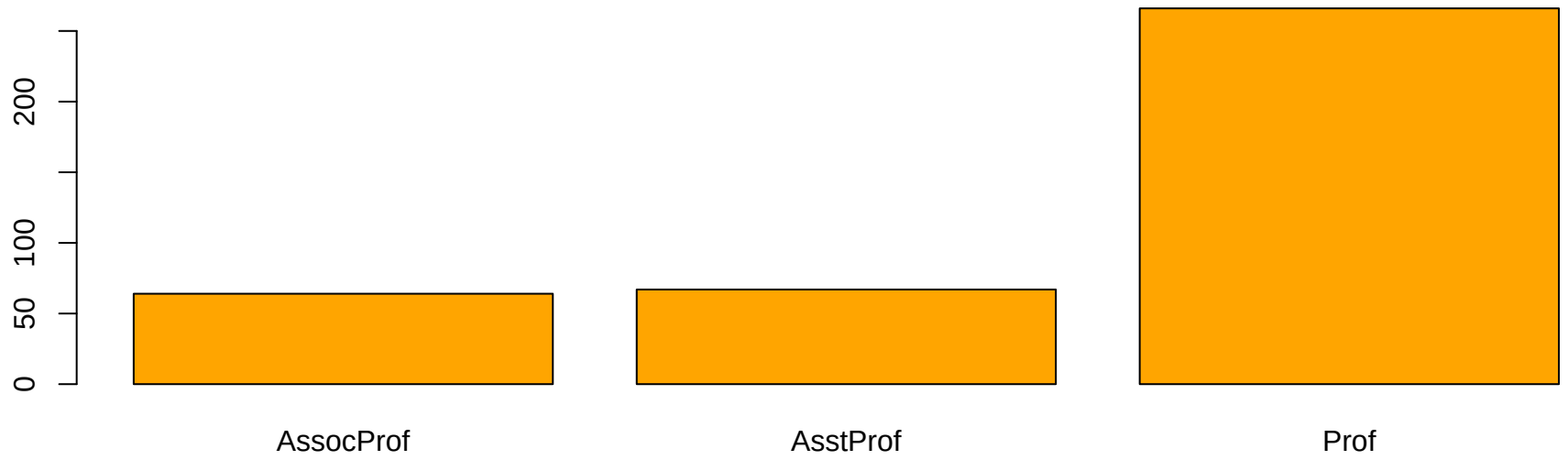
```
tab_discipline <- table(dat$discipline)
tab_discipline
```

```
  A   B
181 216
```

```
tab_sex <- table(dat$sex)
tab_sex
```

สำรวจการแจกแจงของตัวแปร : ตัวแปรจัดประเภท

```
1 barplot(tab_rank, col = "orange")
```



ดำเนินการจัดกระทำข้อมูล

ใส่ label ให้กับ `discipline`

```
dat$discipline <- factor(dat$discipline,
                         levels = c("A","B"),
                         labels = c("Pure Science",
                                    "Applied Science"))
```

สลับ level ของตำแหน่งวิชาการให้เป็นไปตามลำดับ

```
dat <- dat %>%
       mutate(rank = factor(rank,
                            levels = c("AsstProf",
                                       "AssocProf",
                                       "Prof")))
```

สำรวจความสัมพันธ์ระหว่างตัวแปร

การสำรวจความสัมพันธ์ระหว่างตัวแปร

จำแนกตามประเภทของตัวแปร

- ความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระ
- ความสัมพันธ์ระหว่างตัวแปรอิสระ

จำแนกตามประเภทของข้อมูล

- ความสัมพันธ์ระหว่างตัวแปรเชิงปริมาณกับตัวแปรเชิงปริมาณ
- ความสัมพันธ์ระหว่างตัวแปรเชิงปริมาณกับตัวแปรจัดประเภท
- ความสัมพันธ์ระหว่างตัวแปรจัดประเภทกับตัวแปรจัดประเภท

วัตถุประสงค์ลักษณะที่ 1 : เชิงเปรียบเทียบ

วัตถุประสงค์ลักษณะที่ 1

เพื่อเปรียบเทียบค่าเฉลี่ยเงินเดือนของอาจารย์มหาวิทยาลัย ระหว่างกลุ่มอาจารย์ที่มี ตำแหน่งทางวิชาการ สาขาวิชา และเพศแตกต่างกัน

- สำรวจความสัมพันธ์ (วิเคราะห์เปรียบเทียบ) เงินเดือน จำแนกตามตำแหน่งวิชาการ สาขาวิชา และเพศ
- สำรวจความสัมพันธ์ระหว่างตำแหน่งวิชาการ สาขาวิชา และเพศ

ฟังก์ชัน `plot()`

ฟังก์ชัน `plot()` เป็น generic graphic function ซึ่งสามารถใช้ plot แผนภาพที่แตกต่างกันได้ โดยขึ้นกับลักษณะข้อมูลที่นำเข้าในฟังก์ชัน

- ถ้า x และ y เป็นตัวแปรเชิงปริมาณทั้งคู่ ฟังก์ชันจะให้แผนภาพการกระจาย (scatter plot)
- ถ้า x เป็นตัวแปรจัดประเภท และ y เป็นตัวแปรเชิงปริมาณ จะให้ boxplot เชิงเปรียบเทียบ
- ถ้า x เป็นตัวแปรจัดประเภท และ y เป็นตัวแปรจัดประเภท จะให้ mosaic plot

```
1 plot(x, y,
2      col = "black",
3      pch = 16,
4      xlab = "x variable",
5      ylab = "y variable")
```

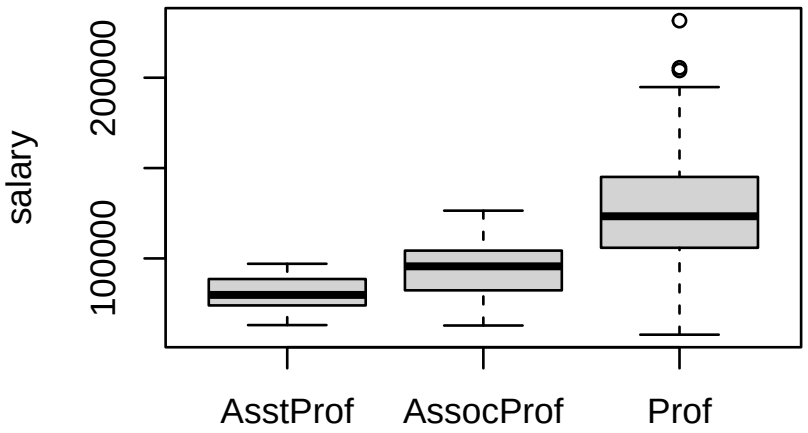
Boxplot เปรียบเทียบเงินเดือนตามตัวแปรอิสระ

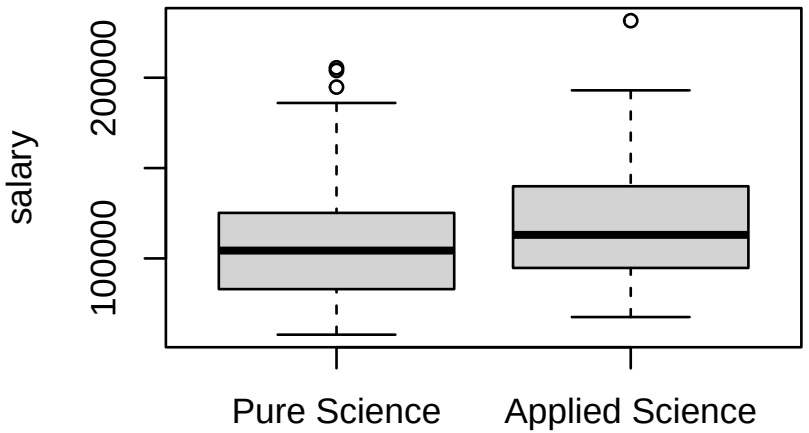
ความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระ

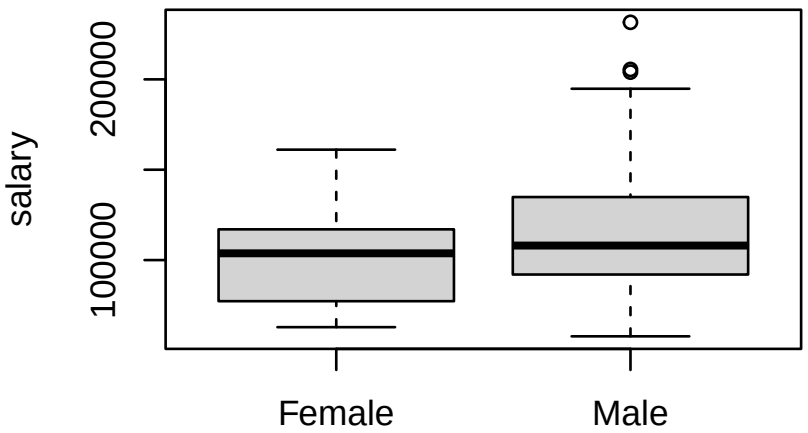
```
par(mfrow = c(2,2), mar=c(3,3,1,1))
plot(x = dat$rank, y = dat$salary,
     xlab = "", ylab = "salary")
plot(x = dat$discipline, y = dat$salary,
    xlab = "", ylab = "salary")
plot(x = dat$sex, y = dat$salary,
    xlab = "", ylab = "salary")
```

Boxplot เปรียบเทียบเงินเดือนตามตัวแปรอิสระ

ความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระ







สถิติพื้นฐานของเงินเดือนจำแนกตาม discipline

ความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระ

```
dat %>% group_by(discipline) %>%
  summarise(mean = mean(salary),
            sd = sd(salary),
            min = min(salary),
            max = max(salary))
```

```
# A tibble: 2 x 5
  discipline          mean     sd   min    max
  <fct>              <dbl>  <dbl> <int>  <int>
1 Pure Science     108548. 30538. 57800 205500
2 Applied Science  118029. 29459. 67559 231545
```

สถิติพื้นฐานของเงินเดือนจำแนกตาม discipline

ความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระ

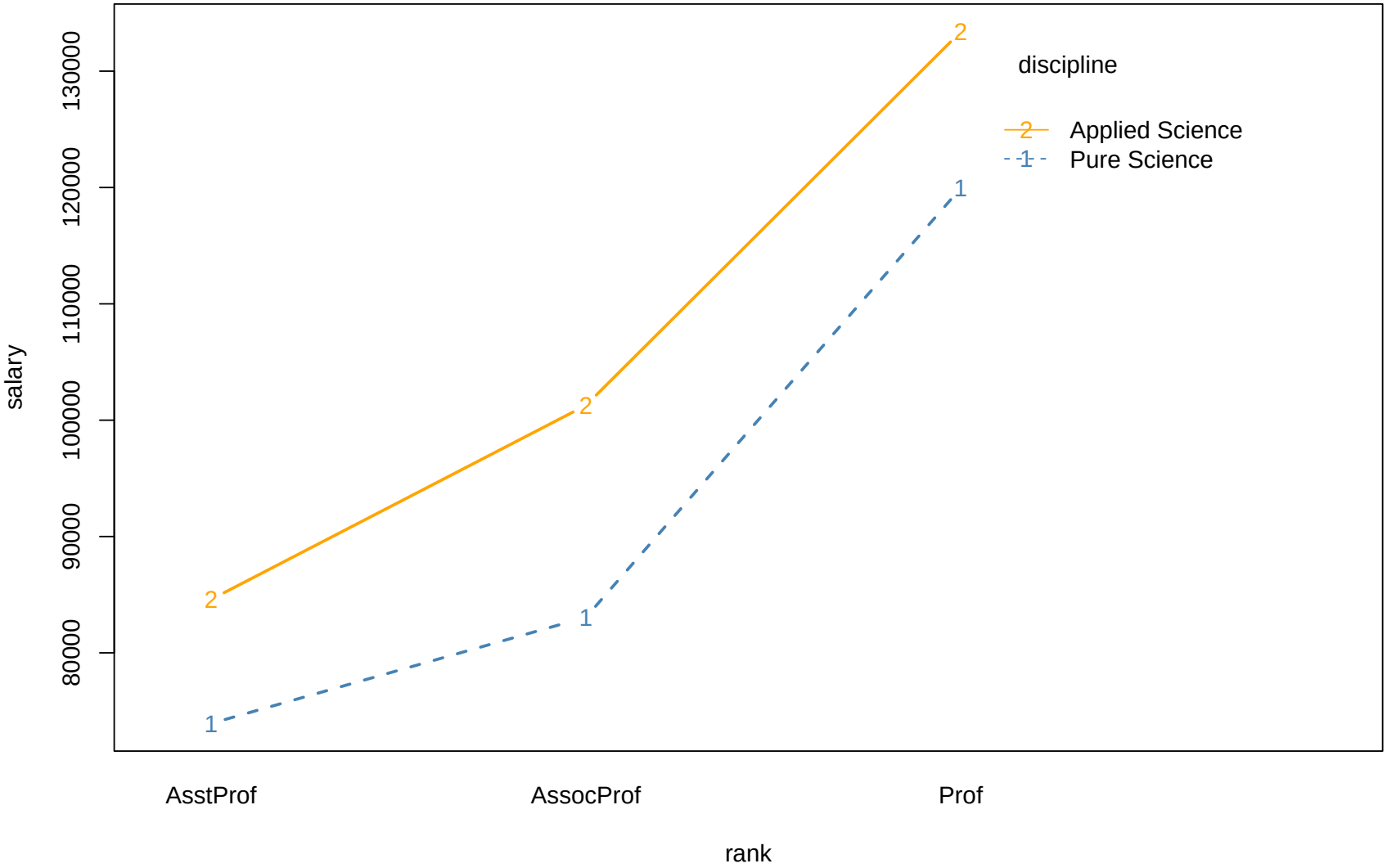
```
dat %>% group_by(rank) %>%
  summarise(mean = mean(salary),
            sd = sd(salary),
            min = min(salary),
            max = max(salary))
```

```
# A tibble: 3 x 5
  rank          mean     sd   min   max
  <fct>        <dbl>  <dbl> <int> <int>
1 AsstProf    80776.  8174. 63100 97032
2 AssocProf   93876. 13832. 62884 126431
3 Prof       126772. 27719. 57800 231545
```

Exploring Interaction Effects (1)

```
interaction.plot(x.factor = dat$rank,
                 trace.factor = dat$discipline,
                 response = dat$salary,
                 xlab = "rank",
                 ylab = "salary",
                 col = c("steelblue","orange"),
                 lwd = 2,
                 trace.label = "discipline",
                 type = "b")
```

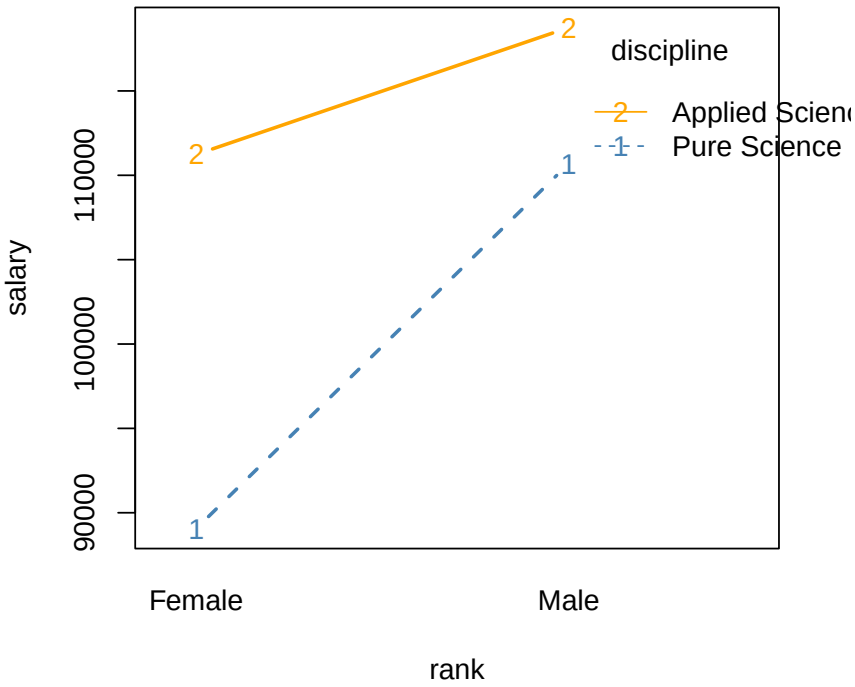
Exploring Interaction Effects (1)

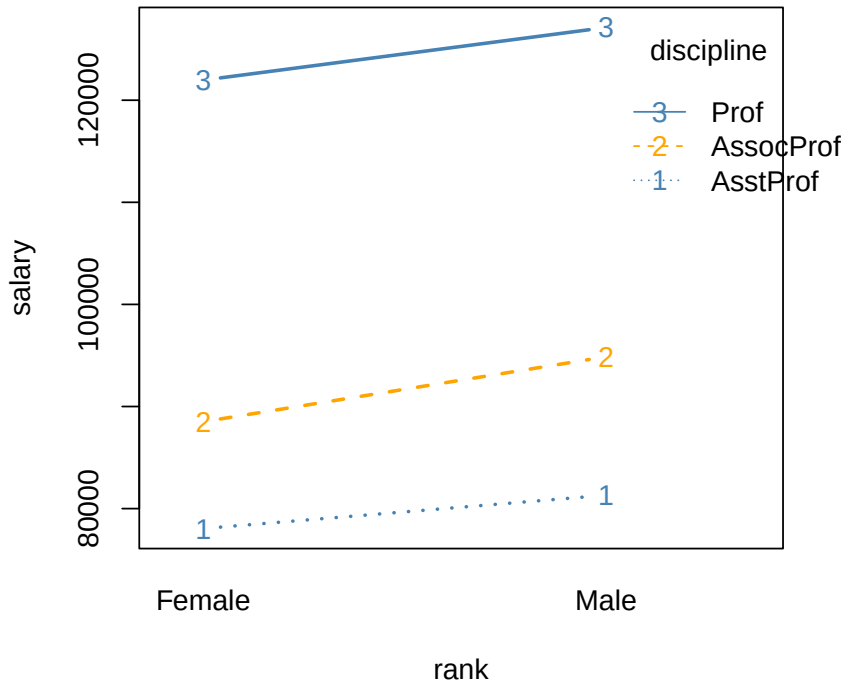


Exploring Interaction Effects (2)

```
par(mfrow=c(1,2))
interaction.plot(x.factor = dat$sex,
                 trace.factor = dat$discipline,
                 response = dat$salary,
                 xlab = "rank",
                 ylab = "salary",
                 col = c("steelblue","orange"),
                 lwd = 2,
                 trace.label = "discipline",
                 type = "b")
interaction.plot(x.factor = dat$sex,
                 trace.factor = dat$rank,
                 response = dat$salary,
                 xlab = "rank",
                 ylab = "salary",
```

Exploring Interaction Effects (2)





สรุปผลการสำรวจ

- เงินเดือนของอาจารย์มหาวิทยาลัย มีแนวโน้มแตกต่างกันได้ตามตำแหน่งวิชาการ สาขาวิชา และเพศ
- มีแนวโน้มว่า จะอิทธิพลปฏิสัมพันธ์ของตำแหน่งวิชาการ และ เพศ ต่อเงินเดือนอาจารย์

Note: ตารางแจกแจงความถี่สองทาง

```
table(dat$rank, dat$discipline)
```

```
           Pure Science Applied Science
  AsstProf           24              43
  AssocProf          26              38
  Prof              131             135
```

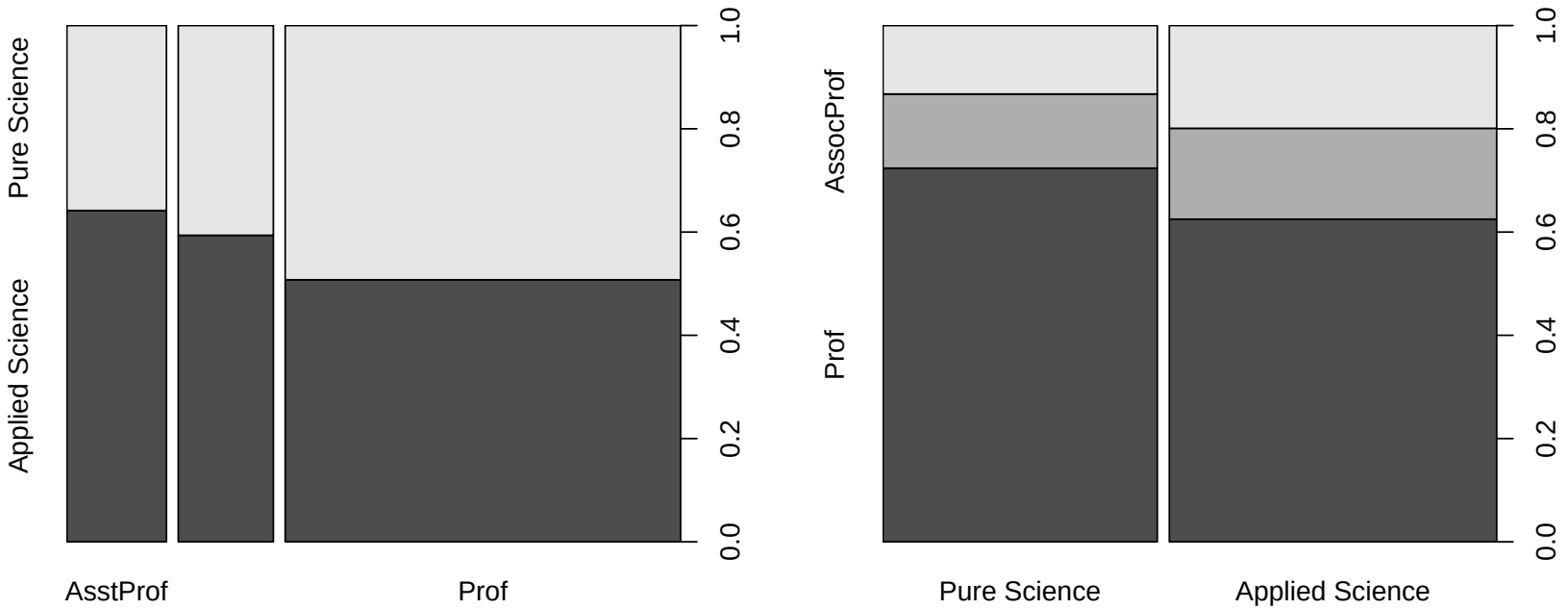
```
table(dat$rank, dat$sex)
```

```
           Female Male
  AsstProf     11   56
  AssocProf    10   54
  Prof         18  248
```

```
table(dat$discipline, dat$sex)
```

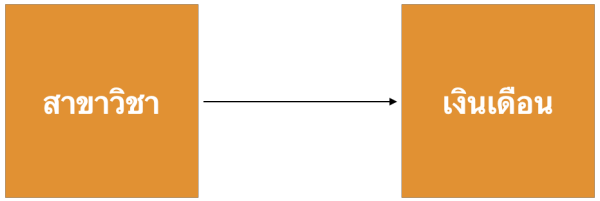
Note: Mosaic plot

```
1 par(mfrow=c(1,2))
2 plot(dat$rank, dat$discipline, xlab = " ", ylab = " ")
3 plot(dat$discipline, dat$rank, xlab = " ", ylab = " ")
```

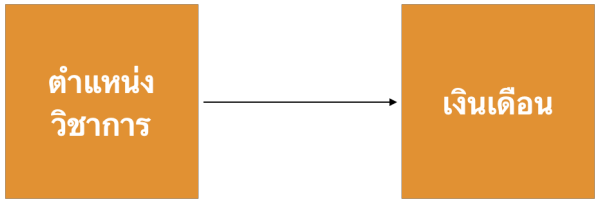


Modelling สำหรับวัตถุประสงค์ลักษณะแรก

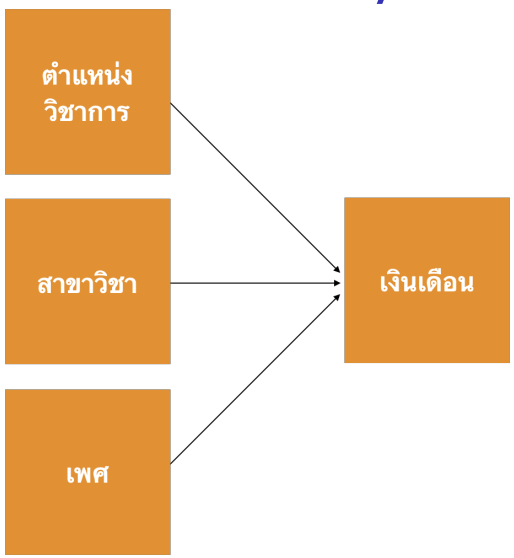
Job 1: independent-sample t-test



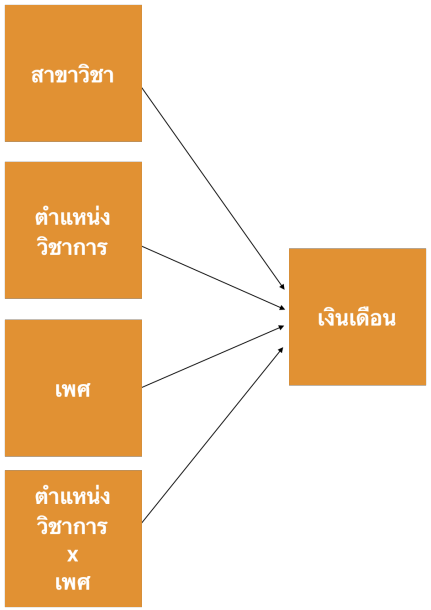
Job 2: One-Way ANOVA



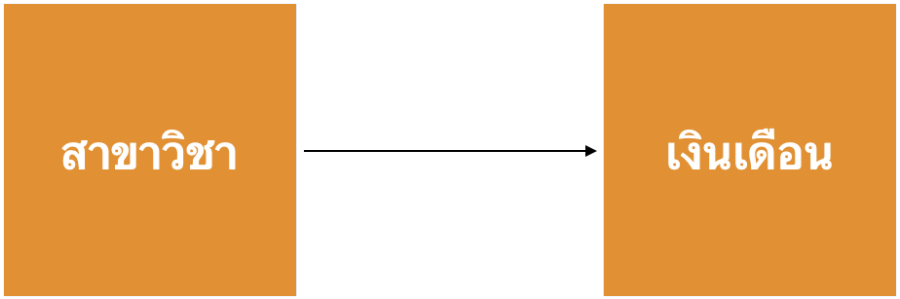
Job 3: three-ways ANOVA



Job 4: Three-Way ANOVA 2



Job1: Independent-sample t-test



สมมุติฐานการทดสอบ

$$H_0 : \mu_{applied} \leq \mu_{pure}$$

$$H_1 : \mu_{applied} > \mu_{pure}$$

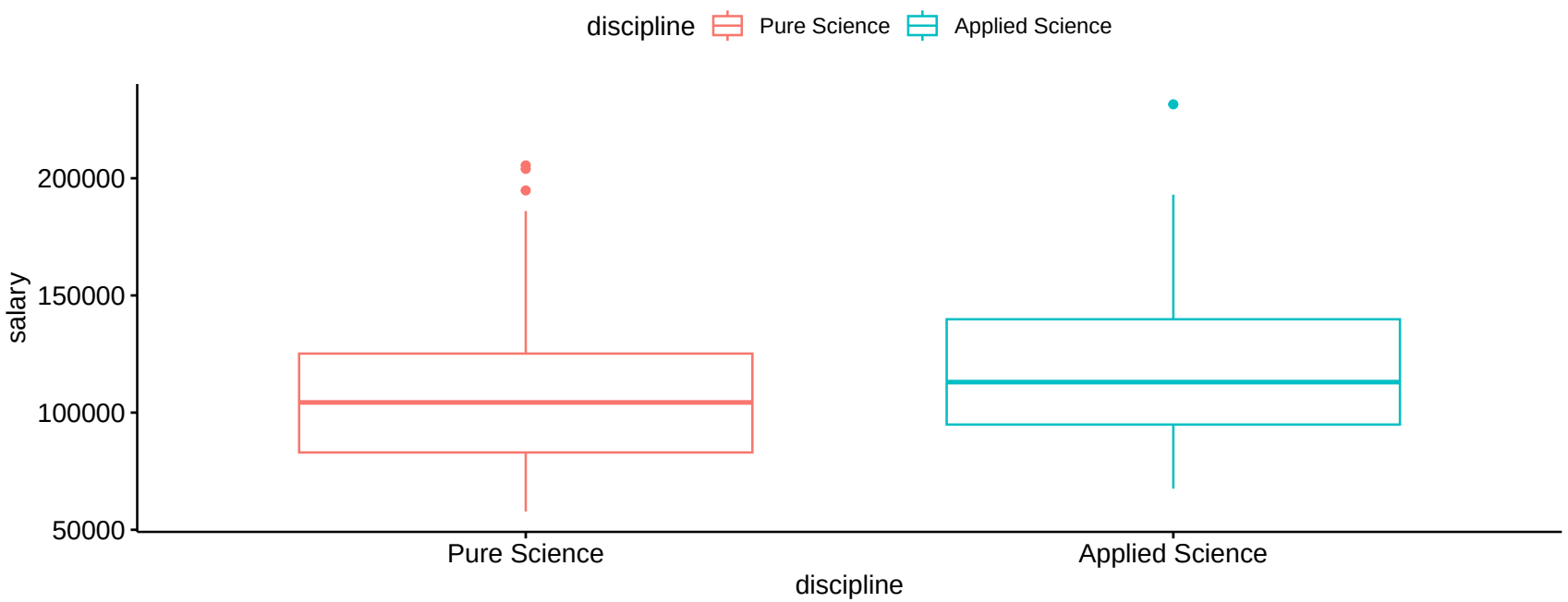
Job1: Independent-sample t-test

```
1 dat %>% group_by(discipline) %>%
2   summarise(mean = mean(salary),
3             sd = sd(salary),
4             min = min(salary),
5             max = max(salary))
```

```
# A tibble: 2 x 5
  discipline          mean    sd   min   max
  <fct>              <dbl> <dbl> <int> <int>
1 Pure Science     108548. 30538. 57800 205500
2 Applied Science 118029. 29459. 67559 231545
```

Job1: Boxplot using ggpubr

```
#install.packages("ggpubr")
library(ggpubr)
ggboxplot(data = dat, x = "discipline", y = "salary",
          color = "discipline")
```



Job1: Independent-sample t-test

ขั้นตอนการวิเคราะห์

Independent sample t-test

ตรวจสอบได้ด้วย Levene's test

$H_0: \sigma^2_{ACH_{PBL}} = \sigma^2_{ACH_{Lec}}$

$H_1: \sigma^2_{ACH_{PBL}} \neq \sigma^2_{ACH_{Lec}}$

Equal variances

Yes. (Don't reject H0 of Levene's test.)

No. (Reject H0 of Levene's test)

Pooled variances t-

$$t^* = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}; df = n_1 + n_2 - 2$$

โดยที่ S_p คือ pooled standard deviation

$$S_p = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$$

Separated variances t-test

$$t^* = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^2}{n_1}\right)^2\left(\frac{1}{n_1 - 1}\right) + \left(\frac{S_2^2}{n_2}\right)\left(\frac{1}{n_2 - 1}\right)}$$

Job1: Levene's Test for equality of variances

```
#install.packages("car")
library(car)
leveneTest(salary ~ discipline, data = dat)
```

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1  0.0458 0.8306
      395
```

Job1: Syntax for t-test

```
# two-sided test
t.test(salary ~ discipline ,data=dat,
       var.equal = TRUE)
# one-sided test
t.test(salary ~ discipline, data=dat,
       var.equal = FALSE,
       alternative = "greater")
```

Job1: Independent-sample t-test output

```r
t.test(salary ~ discipline, data=dat,
       var.equal = FALSE,
       alternative = "less")
```

```
    Welch Two Sample t-test

data:  salary by discipline
t = -3.1306, df = 377.83, p-value = 0.00094
alternative hypothesis: true difference in means between gr
95 percent confidence interval:
      -Inf -4487.034
sample estimates:
   mean in group Pure Science mean in group Applied Science
                     108548.4                      118028.7
```

Job1: Assumptions check

independent sample t-test มีข้อตกลงเบื้องต้นที่สำคัญดังนี้

- Independence
- Normality
- Homogeneity of variances

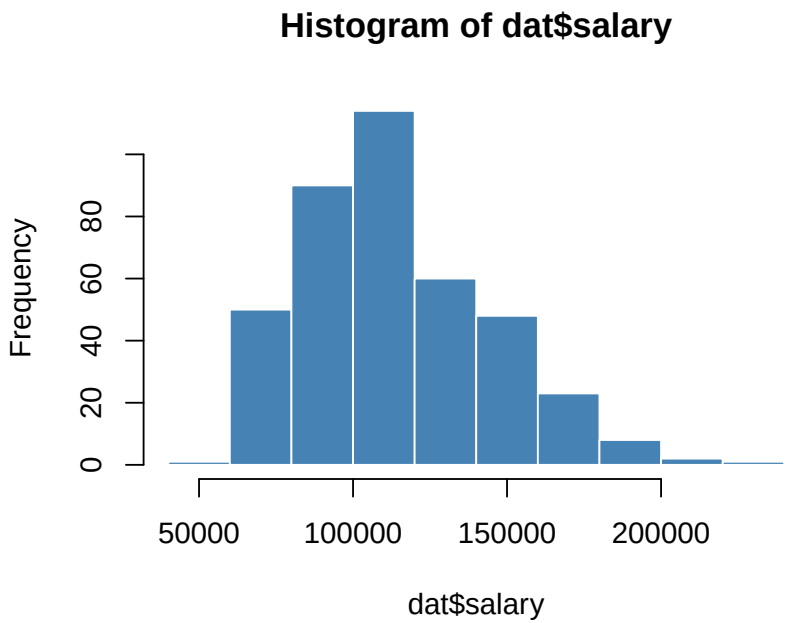
Job1: Assumptions check

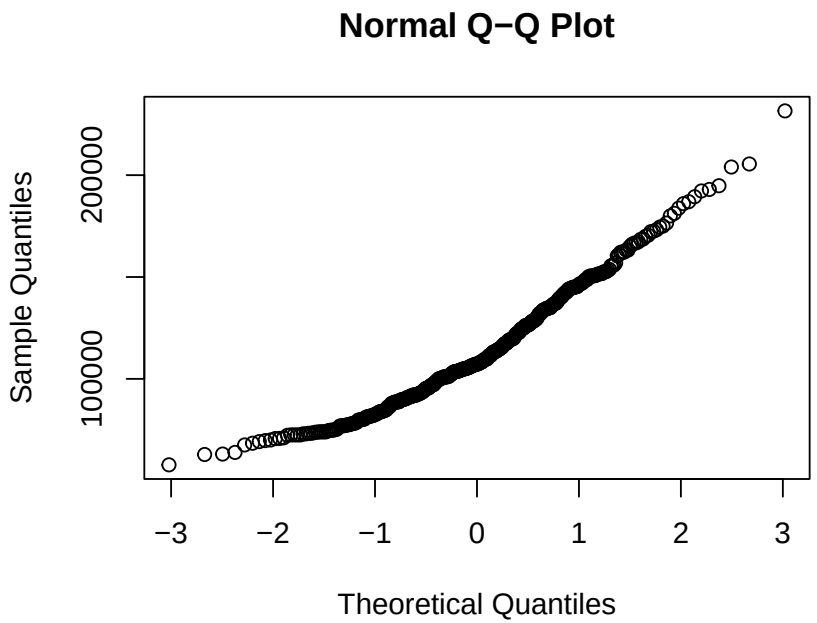
```
1 shapiro.test(dat$salary)
```

```
    Shapiro-Wilk normality test

data:  dat$salary
W = 0.95988, p-value = 6.076e-09
```

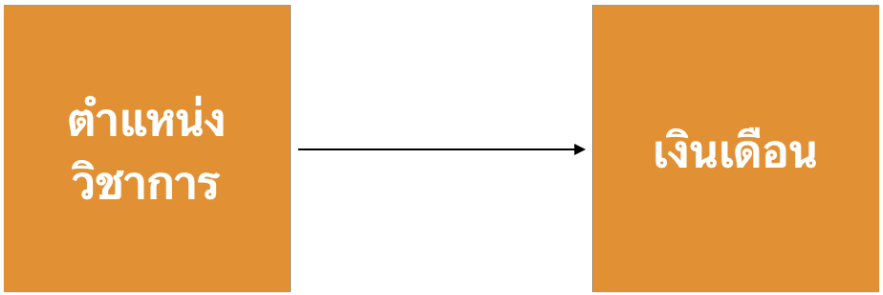
Job1: Assumptions check

```
1 par(mfrow=c(1,2))
2 hist(dat$salary, col="steelblue", border = "white")
3 qqnorm(dat$salary)
```





Job2: Independent-sample F-test (One-Way ANOVA)

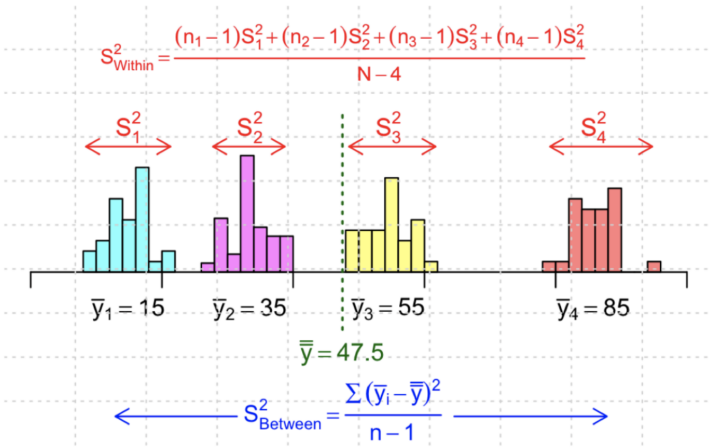



```
dat %>% group_by(rank) %>%
    summarise(mean = mean(salary),
              sd = sd(salary),
              min = min(salary),
              max = max(salary))%>%
  arrange(desc(mean))
```

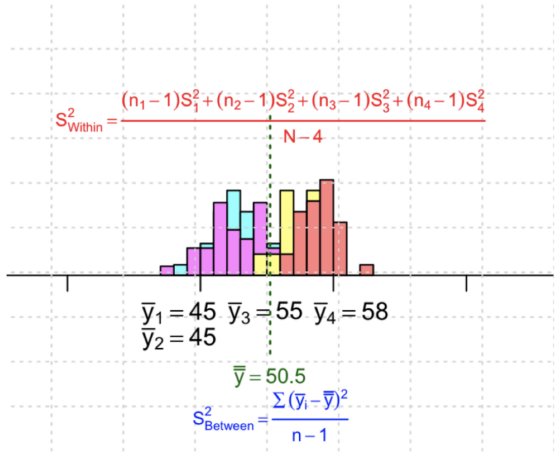
Job2: Independent-sample F-test (One-Way ANOVA)

```
# A tibble: 3 x 5
  rank          mean     sd   min    max
  <fct>        <dbl>  <dbl> <int>  <int>
1 Prof      126772. 27719. 57800 231545
2 AssocProf  93876. 13832. 62884 126431
3 AsstProf   80776.  8174. 63100  97032
```

Job2: Independent-sample F-test (One-Way ANOVA)



Note: Between Group Variance > Within Group Variance



Note: Between Group Variance ≈ Within Group Variance

ที่มา : สิวะโชติ ศรีสุทธิยากร (2560)

Job2: Calculate F-test for Overall hypothesis

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_1 : not\ H_0$$

```
# calculate analysis of variance
fit <- aov(salary ~ rank, data= dat)
summary(fit)
```

```
             Df    Sum Sq   Mean Sq F value Pr(>F)
rank          2 1.432e+11 7.162e+10   128.2 <2e-16 ***
Residuals   394 2.201e+11 5.586e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

ผลการวิเคราะห์ข้างต้นสรุปว่า ...

Job2: Multiple comparison

การทดสอบเพื่อเปรียบเทียบค่าเฉลี่ยรายคู่มีหลายวิธีการ

- `TukeyHSD()`
- `pairwise.t.test()`
- `ScheffeTest()`

Job2: Calculate Tukey HSD

```
1 TukeyHSD(fit)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = salary ~ rank, data = dat)

$rank
                          diff       lwr      upr     p adj
AssocProf-AsstProf 13100.45  3382.195 22818.71 0.0046514
Prof-AsstProf      45996.12 38395.941 53596.31 0.0000000
Prof-AssocProf     32895.67 25154.507 40636.84 0.0000000
```

Job2: Calculate Bonferroni

```
pairwise.t.test(x = dat$salary, g = dat$rank,
                pool.sa = TRUE, p.adjust.method = "bonferro
```

```
    Pairwise comparisons using t tests with pooled SD

data:  dat$salary and dat$rank

          AsstProf AssocProf
AssocProf 0.0049   -
Prof      <2e-16   <2e-16

P value adjustment method: bonferroni
```

```
# c("holm", "hochberg", "hommel", "bonferroni", "BH", "BY",
#   "fdr", "none")
```

Job2: Calculate Scheffe's test

```
1 #install.packages("DescTools")
2 library(DescTools)
3 ScheffeTest(fit)
```

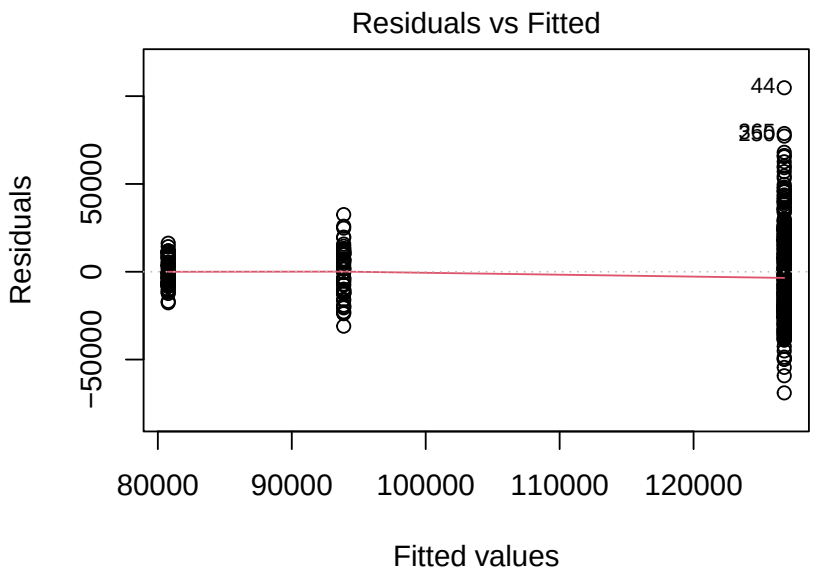
```
  Posthoc multiple comparisons of means: Scheffe Test
    95% family-wise confidence level

$rank
                          diff    lwr.ci   upr.ci   pval
AssocProf-AsstProf  13100.45  2950.615 23250.29 0.0070 **
Prof-AsstProf       45996.12 38058.423 53933.83 <2e-16 ***
Prof-AssocProf      32895.67 24810.728 40980.62 <2e-16 ***

---
```

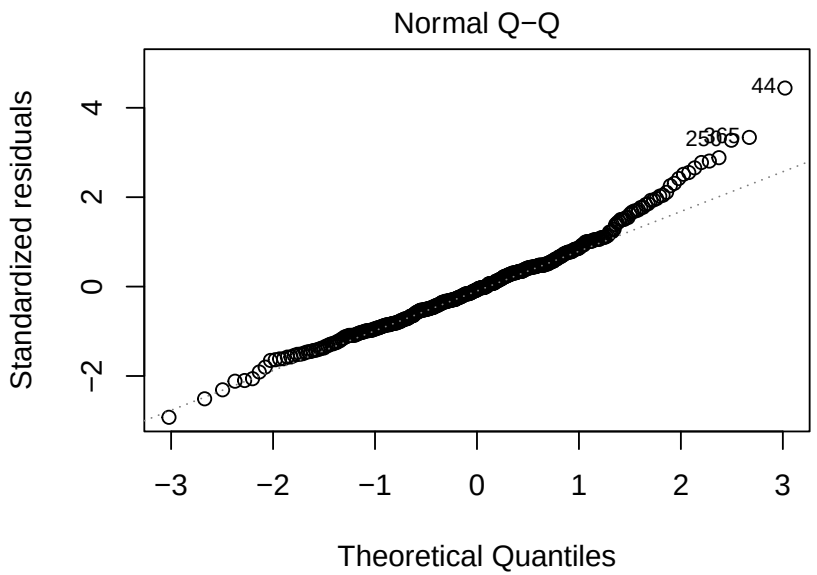
Job2: Assumption Checking

- Normality
- Homogeneity of variances

Job2: Assumption Checking

```
1 par(mfrow=c(1,2))
2 plot(fit, 1:2)
```



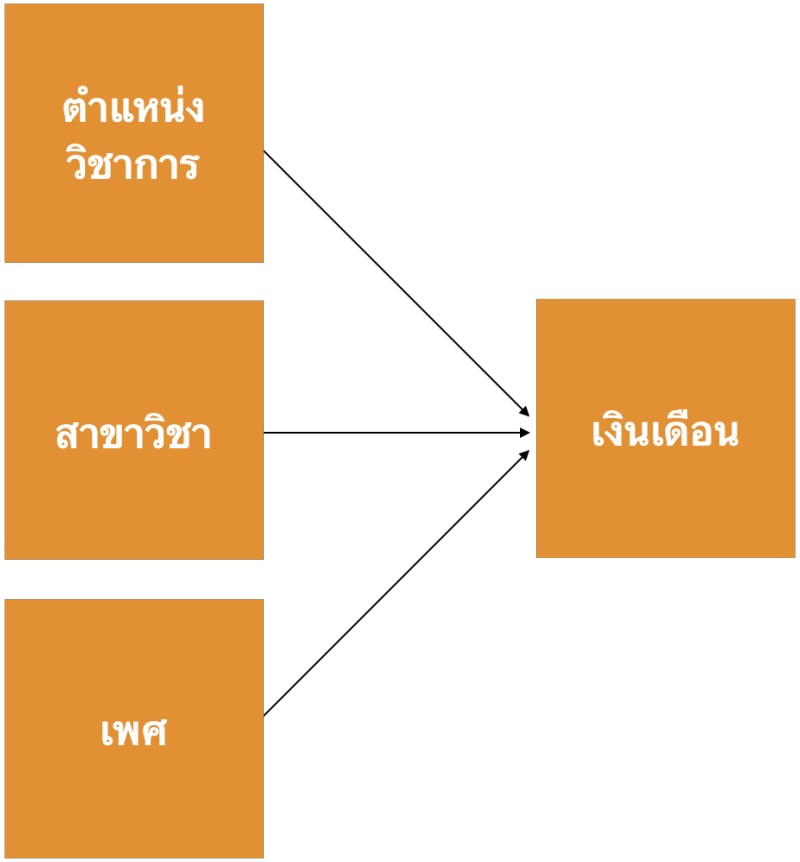


Job2: Welch's F-test

```r
oneway.test(salary ~ rank, data = dat, var.equal = F)
```

```
        One-way analysis of means (not assuming equal variances

data:  salary and rank
F = 271.44, num df = 2.00, denom df = 177.19, p-value < 2.2
```

Job3: Three-Way ANOVA



Job3: Summary Stat

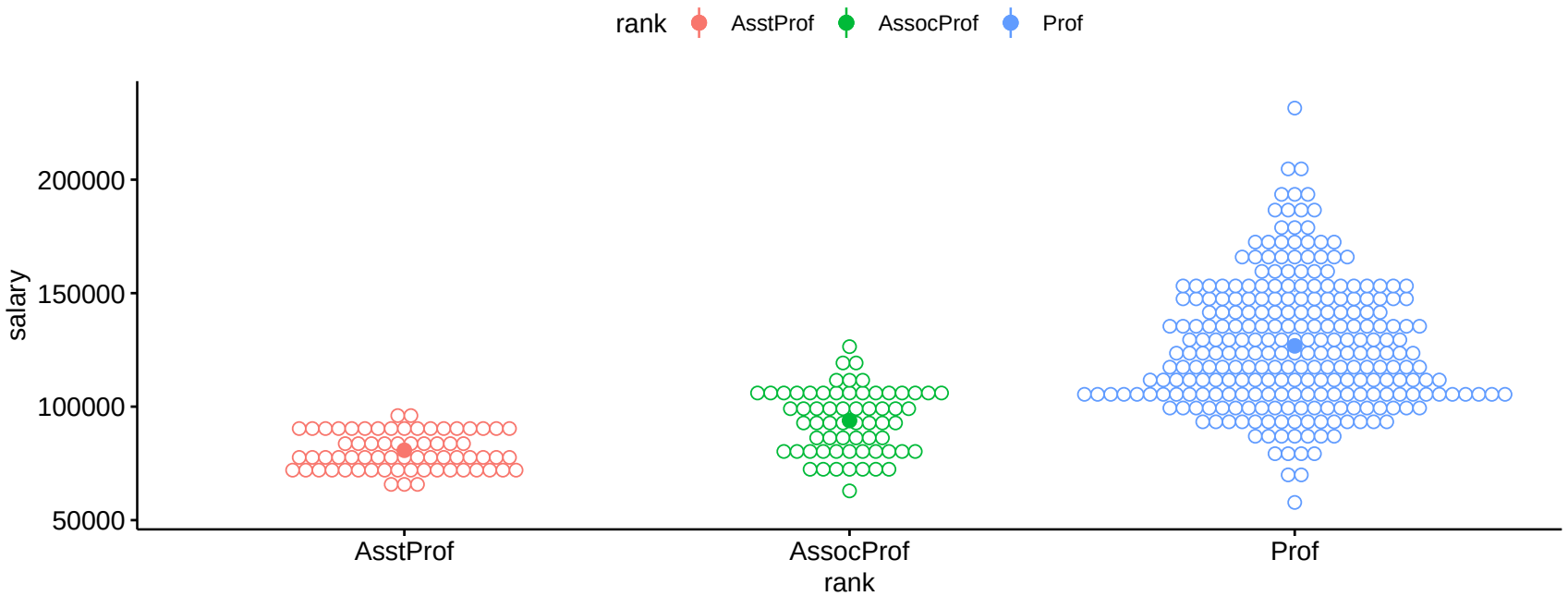
```
dat %>% gather(rank, discipline, sex,
                key = "ind",
                value = "value")%>%
  group_by(ind, value) %>%
  summarise(mean = mean(salary),
            sd = sd(salary))
```

Job3: Summary Stat

```
# A tibble: 7 x 4
# Groups:   ind [3]
  ind        value              mean     sd
  <chr>      <chr>             <dbl>  <dbl>
1 discipline Applied Science 118029. 29459.
2 discipline Pure Science    108548. 30538.
3 rank       AssocProf        93876. 13832.
4 rank       AsstProf         80776.  8174.
5 rank       Prof            126772. 27719.
6 sex        Female          101002. 25952.
7 sex        Male            115090. 30437.
```

Job3: Summary Stat

```
ggdotplot(data = dat, x = "rank", y = "salary",
          add = c("mean_se","dotplot"),
          col = "rank",
          binwidth = 0.2)
```



Job3: Calculate F-test for Three-Way ANOVA

```r
fit3 <- aov(salary ~ rank + discipline + sex ,
            data = dat)
summary(fit3)
```

```
               Df    Sum Sq   Mean Sq F value   Pr(>F)
rank            2 1.432e+11 7.162e+10 139.707  < 2e-16 ***
discipline      1 1.843e+10 1.843e+10  35.953 4.59e-09 ***
sex             1 6.941e+08 6.941e+08   1.354    0.245
Residuals     392 2.009e+11 5.126e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Job3: Calculate Multiple Comparison

```
1 TukeyHSD(fit3)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = salary ~ rank + discipline + sex, data =

$rank
                         diff       lwr      upr     p adj
AssocProf-AsstProf 13100.45  3790.209 22410.70 0.0029189
Prof-AsstProf      45996.12 38715.029 53277.22 0.0000000
Prof-AssocProf     32895.67 25479.514 40311.83 0.0000000

$discipline
                                diff       lwr      upr p a
```

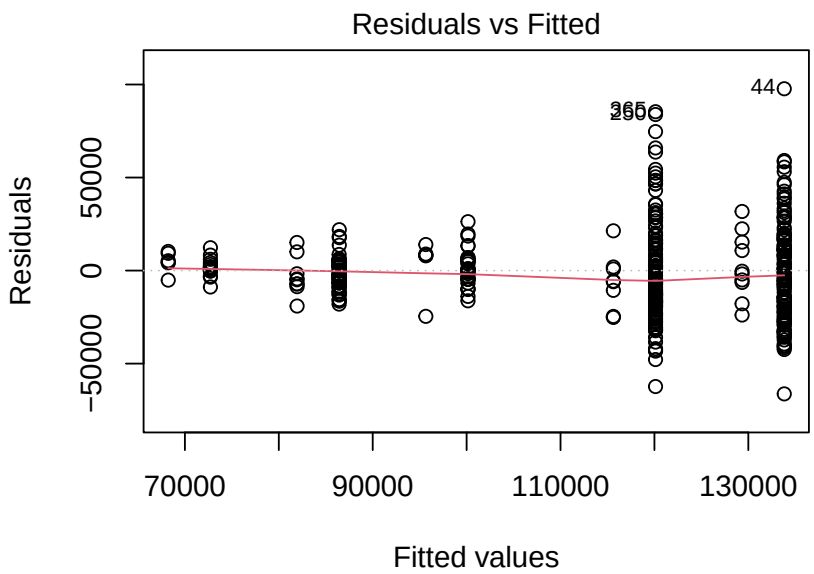
Job3: Calculate Multiple Comparison

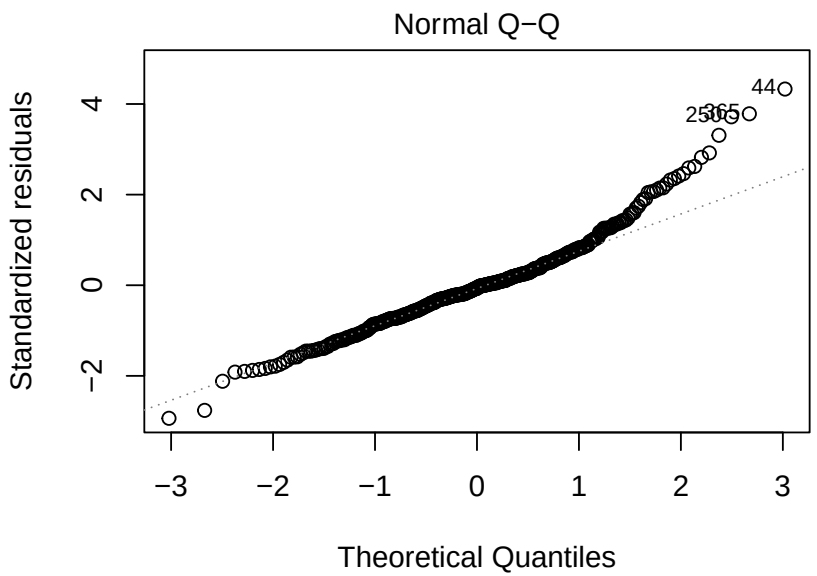
```
1 ScheffeTest(fit3)
```

```
  Posthoc multiple comparisons of means: Scheffe Test
    95% family-wise confidence level

$rank
                          diff     lwr.ci   upr.ci   pval
AssocProf-AsstProf 13100.45   852.6023 25348.30 0.0285 *
Prof-AsstProf      45996.12 36417.6681 55574.58 <2e-16 ***
Prof-AssocProf     32895.67 23139.5373 42651.81 <2e-16 ***

$discipline
                                      diff   lwr.ci   upr.ci
Applied Science-Pure Science 13599.82 6538.587 20661.06 7.1
```

Job3: Assumptions Checking

```
1 par(mfrow=c(1,2))
2 plot(fit3, 1:2)
```



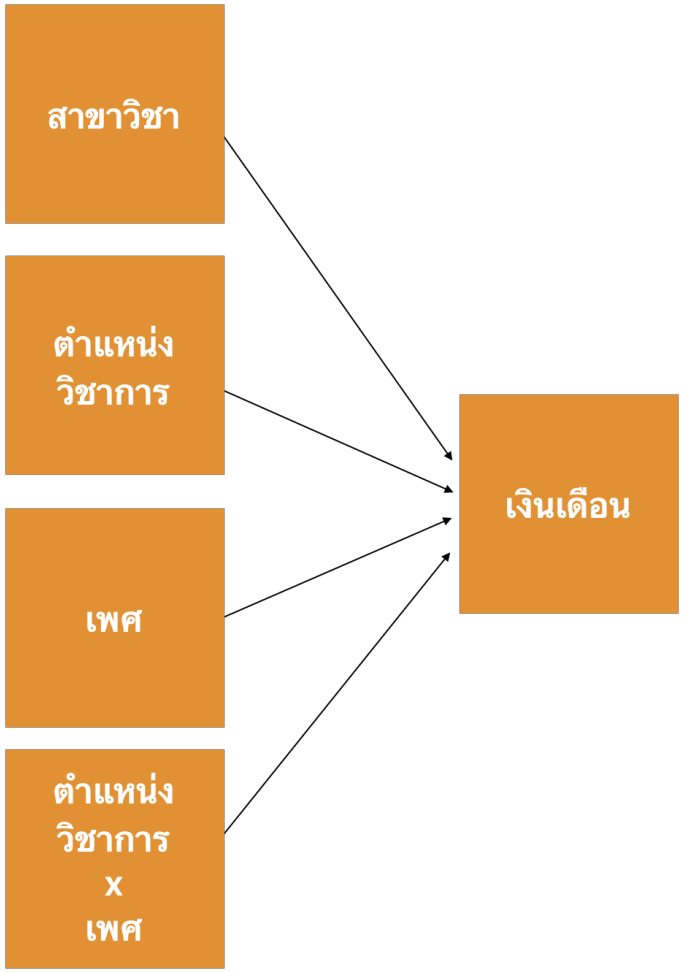


Job3: Assumptions Checking

```
1 leveneTest(salary ~ rank*discipline*sex , data = dat)
```

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group  11   9.047 2.064e-14 ***
      385
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Job4: Three-Way ANOVA 2



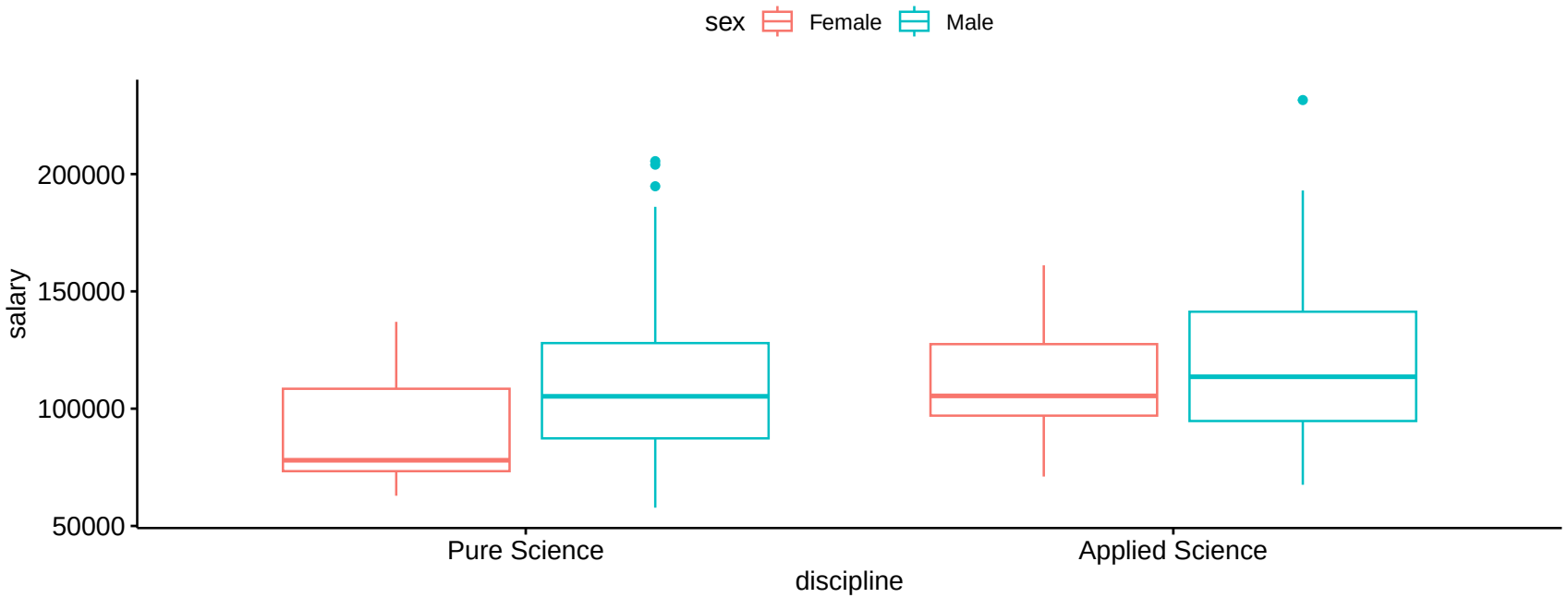
Job4: Summary Stat

```
1 dat %>% group_by(rank, sex) %>%
2   summarise(mean = mean(salary),
3             sd = sd(salary))
```

```
# A tibble: 6 x 4
# Groups:   rank [3]
  rank      sex       mean     sd
  <fct>     <fct>    <dbl>  <dbl>
1 AsstProf  Female  78050.  9372.
2 AsstProf  Male    81311.  7901.
3 AssocProf Female  88513. 17965.
4 AssocProf Male    94870. 12891.
5 Prof      Female 121968. 19620.
6 Prof      Male   127121. 28214.
```

Job4: Boxplot

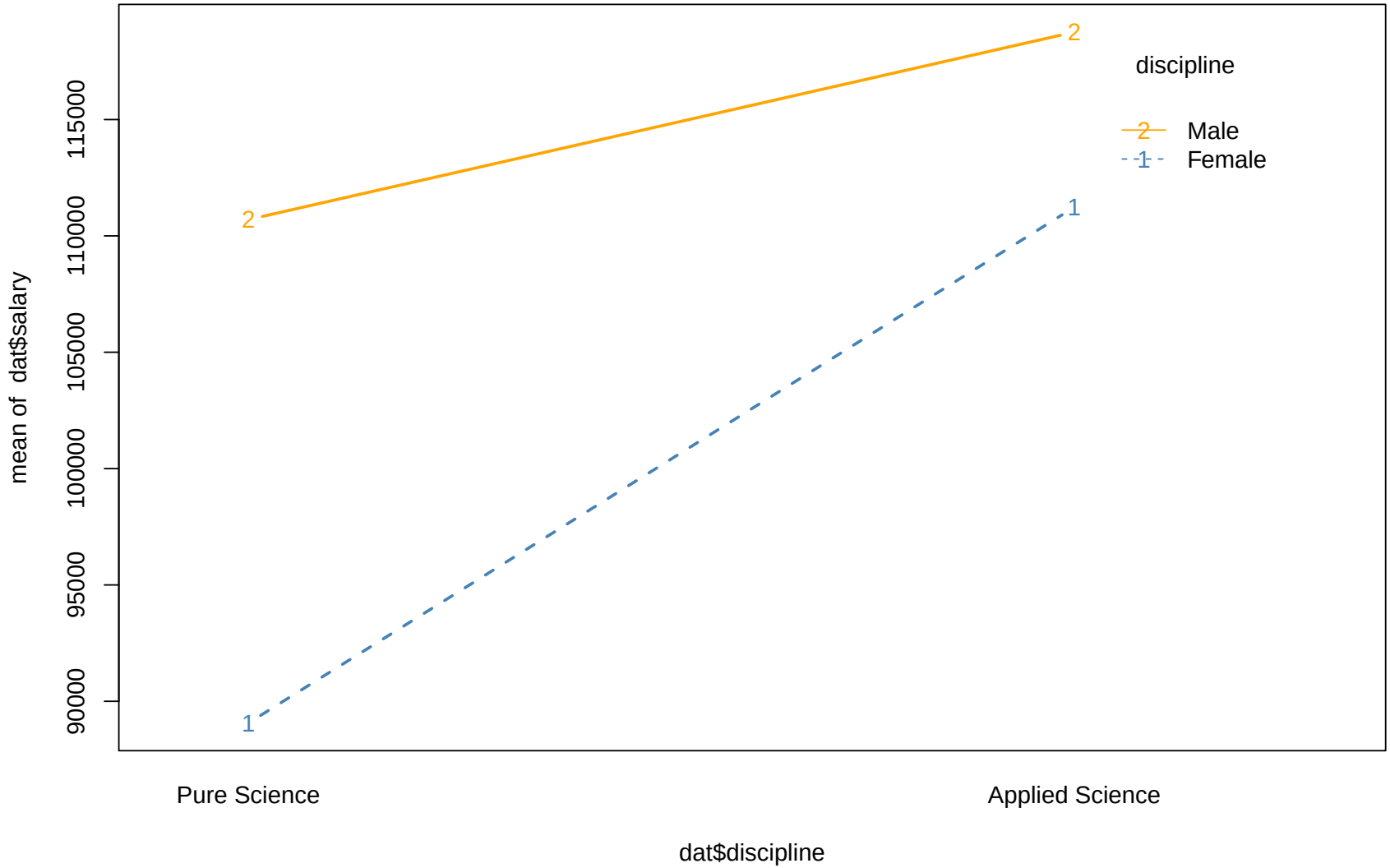
```
1 ggboxplot(data = dat, x = "discipline", y = "salary",
2           col = "sex")
```



Job4: interaction plot

```
interaction.plot(x.factor = dat$discipline,
                 trace.factor = dat$sex,
                 response = dat$salary,
                 col = c("steelblue","orange"),
                 lwd = 2,
                 trace.label = "discipline",
                 type = "b")
```

Job4: interaction plot



Job4: calculate ANOVA with interaction

```
fit_job4<-aov(salary ~ rank + discipline + sex + discipline
              data = dat)
summary(fit_job4)
```

```
                Df    Sum Sq   Mean Sq F value   Pr(>F)
rank             2 1.432e+11 7.162e+10 139.629  < 2e-16 ***
discipline       1 1.843e+10 1.843e+10  35.933 4.65e-09 ***
sex              1 6.941e+08 6.941e+08   1.353    0.245
discipline:sex   1 4.001e+08 4.001e+08   0.780    0.378
Residuals      391 2.005e+11 5.129e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Job4: Simple Effect Analysis

```
1 library(phia)
2 testInteractions(fit_job4,
3                  fixed = "discipline",
4                  across = "sex",
5                  adjustment = "bonferroni")
```

```
F Test:
P-value adjustment method: bonferroni
                    Value  Df  Sum of Sq      F Pr(>F)
   Pure Science -8173.9    1 1.0612e+09 2.0691 0.3022
Applied Science -1390.0    1 3.6372e+07 0.0709 1.0000
Residuals                391 2.0054e+11
```

Job4: Simple Effect Analysis

```
1 testInteractions(fit_job4,
2                  fixed = "sex",
3                  across = "discipline",
4                  adjustment = "bonferroni",
5                  pairwise = "discipline")
```

```
F Test:
P-value adjustment method: bonferroni
                                              Value  Df  Sum of Sc
Pure Science-Applied Science : Female -19812    1 3.7968e+09
Pure Science-Applied Science :   Male -13028    1 1.4839e+10
Residuals                                      391 2.0054e+11

Pure Science-Applied Science : Female *
Pure Science-Applied Science :   Male ***
```