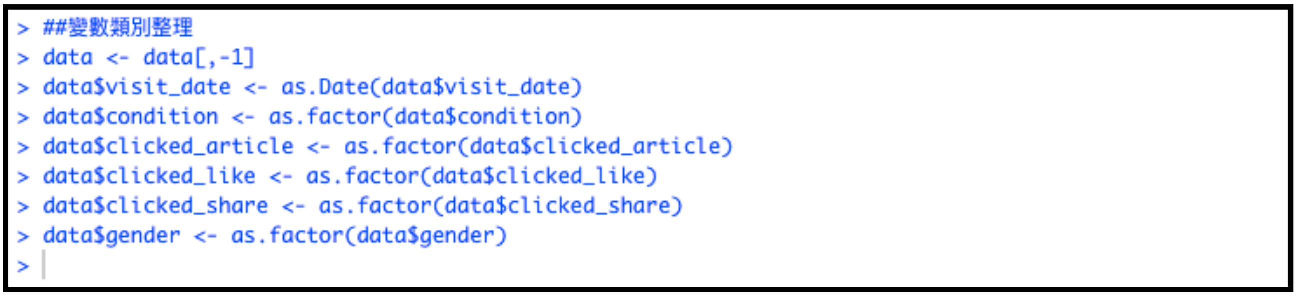
SAS/R商業資料分析作業六

107508006 歐西四 陳葳芃

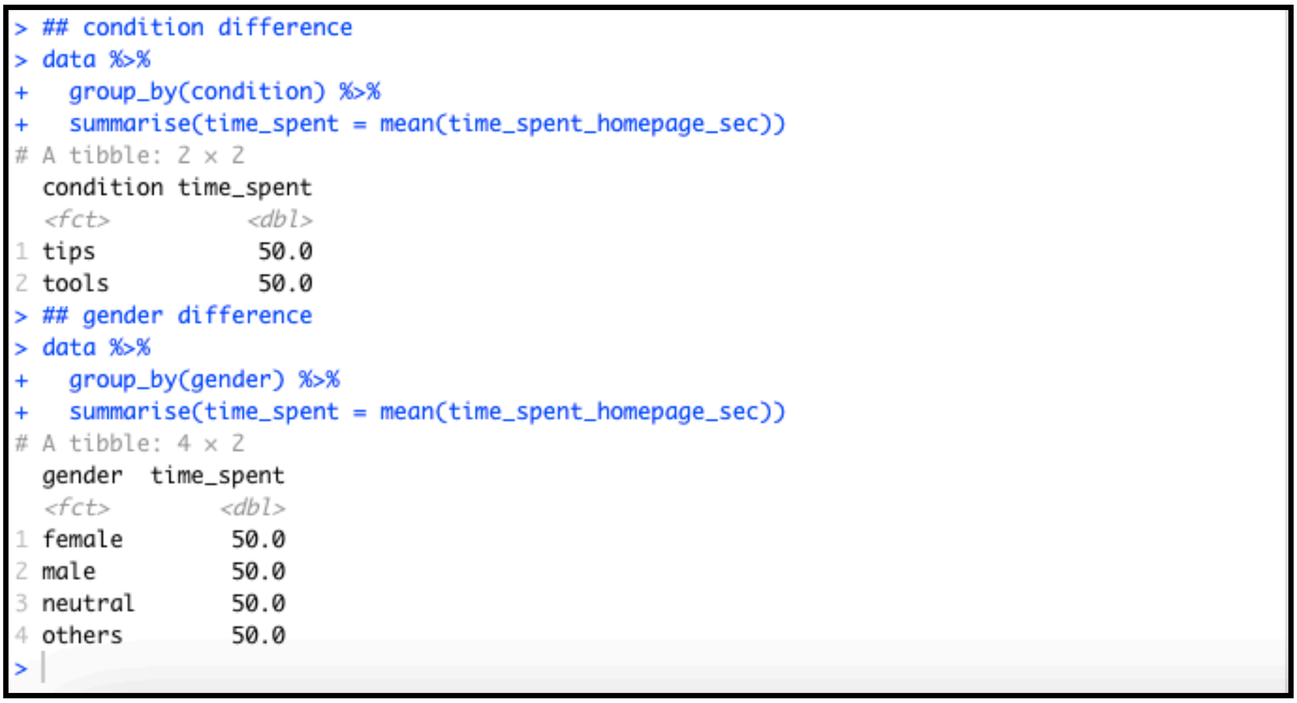
1. 某網紅想分析他Facebook上寫的文章。他的文章分為兩種(condition)：建議(tips)和工具(tools)。利用A/B Testing課程所教的，畫圖及用檢定方法，幫助網紅分析他的粉絲喜歡哪種文章，以後該網紅應該多寫哪種文章來增加觸擊率。（可自行決定你要分析的面相，如按讚率或分享率等。）

Ans:



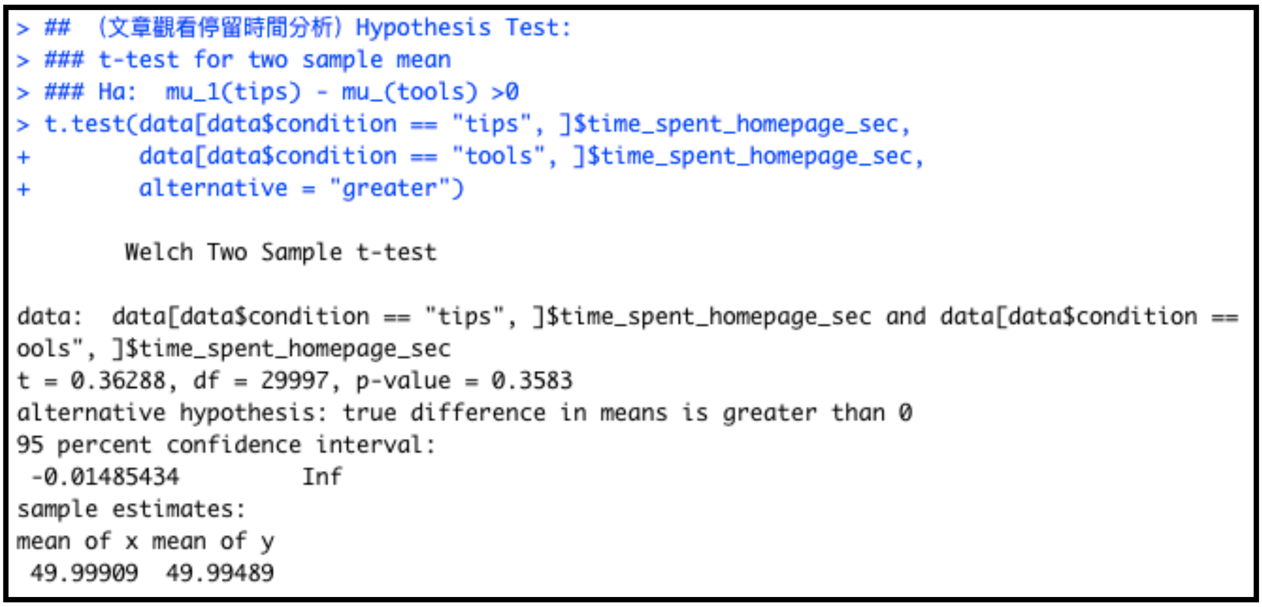
```
> ##變數類別整理
> data <- data[,-1]
> data$visit_date <- as.Date(data$visit_date)
> data$condition <- as.factor(data$condition)
> data$clicked_article <- as.factor(data$clicked_article)
> data$clicked_like <- as.factor(data$clicked_like)
> data$clicked_share <- as.factor(data$clicked_share)
> data$gender <- as.factor(data$gender)
> |
```

>>變數轉換



```
> ## condition difference
> data %>%
+   group_by(condition) %>%
+   summarise(time_spent = mean(time_spent_homepage_sec))
# A tibble: 2 x 2
  condition time_spent
  <fct>          <dbl>
1 tips            50.0
2 tools           50.0
> ## gender difference
> data %>%
+   group_by(gender) %>%
+   summarise(time_spent = mean(time_spent_homepage_sec))
# A tibble: 4 x 2
  gender   time_spent
  <fct>         <dbl>
1 female         50.0
2 male           50.0
3 neutral        50.0
4 others         50.0
> |
```

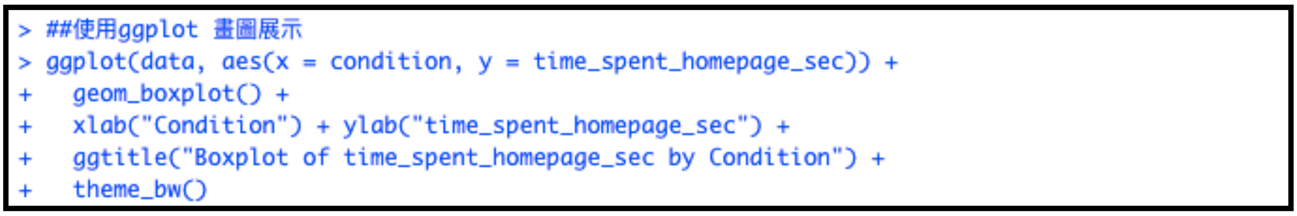
>>分析：從此分析可看到在網頁兩種文章停留的平均時間上，數字上來說沒有明顯差異，後續則建立一假設檢定來判別在統計上是否真的沒有明顯差異。



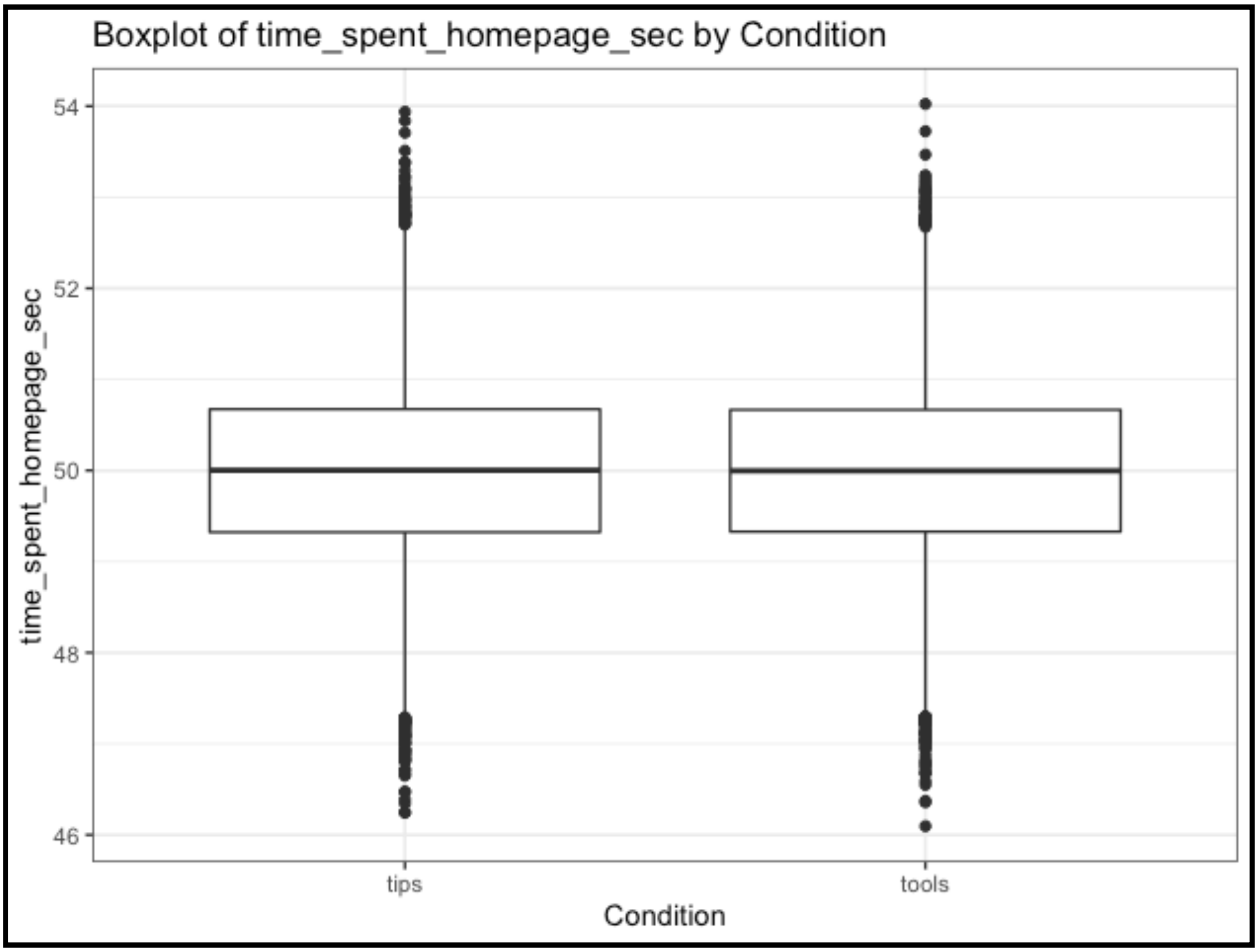
```
> ##  (文章觀看停留時間分析) Hypothesis Test:
> ### t-test for two sample mean
> ### Ha:  mu_1(tips) - mu_(tools) >0
> t.test(data[data$condition == "tips", ]$time_spent_homepage_sec,
+        data[data$condition == "tools", ]$time_spent_homepage_sec,
+        alternative = "greater")

        Welch Two Sample t-test

data:  data[data$condition == "tips", ]$time_spent_homepage_sec and data[data$condition ==
ools", ]$time_spent_homepage_sec
t = 0.36288, df = 29997, p-value = 0.3583
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.01485434         Inf
sample estimates:
mean of x mean of y 
 49.99909  49.99489 
```

>>作法：因此我們使用假設檢定單尾t檢定來判別，兩者是否真的在統計上的意義來說是相等的；而從檢定結果我們可以看到p-value的結果我們可以宣稱在alpha=0.05的level我們不拒絕虛無假設。同時，在統計上的意義也顯示其粉絲在網頁上停留的平均時間上沒有隨著寫作種類的不同而有顯著差異。

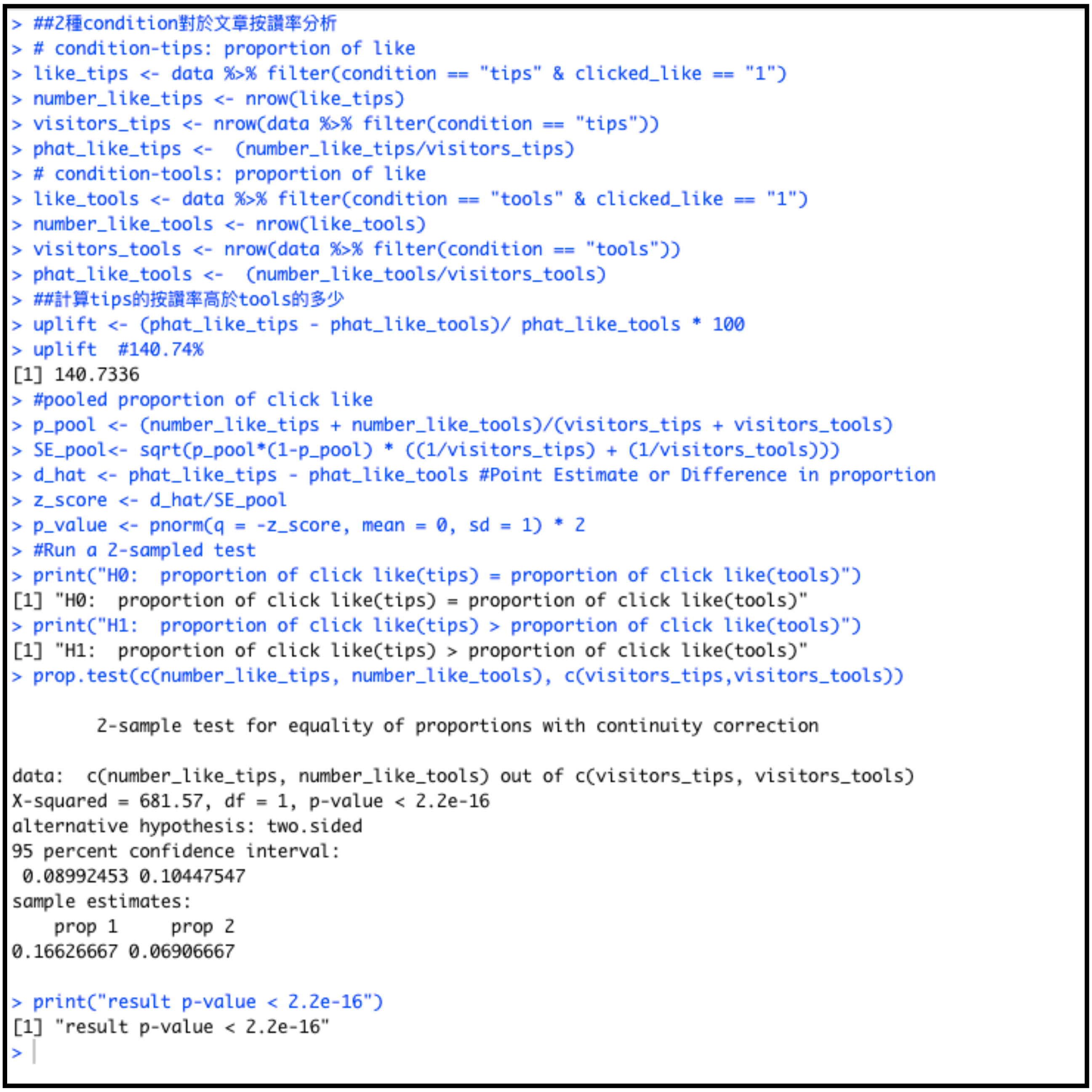
>>分析：根據檢定結果，若想要提升讀者/粉絲在網頁上停留的平均時間，可能可以嘗試其他種寫作方式、或考量其他因素（ex: 發文時間、提升內容品質......等。）



```
> ##使用ggplot 畫圖展示
> ggplot(data, aes(x = condition, y = time_spent_homepage_sec)) +
+   geom_boxplot() +
+   xlab("Condition") + ylab("time_spent_homepage_sec") +
+   ggtitle("Boxplot of time_spent_homepage_sec by Condition") +
+   theme_bw()
```



>>分析建議：而我們也可以用ggplot繪出boxplot來檢視資料情形，同樣也可以看出2種文章沒有顯著差異。



```
> ##2種condition對於文章按讚率分析
> # condition-tips: proportion of like
> like_tips <- data %>% filter(condition == "tips" & clicked_like == "1")
> number_like_tips <- nrow(like_tips)
> visitors_tips <- nrow(data %>% filter(condition == "tips"))
> phat_like_tips <-  (number_like_tips/visitors_tips)
> # condition-tools: proportion of like
> like_tools <- data %>% filter(condition == "tools" & clicked_like == "1")
> number_like_tools <- nrow(like_tools)
> visitors_tools <- nrow(data %>% filter(condition == "tools"))
> phat_like_tools <-  (number_like_tools/visitors_tools)
> ##計算tips的按讚率高於tools的多少
> uplift <- (phat_like_tips - phat_like_tools)/ phat_like_tools * 100
> uplift  #140.74%
[1] 140.7336
> #pooled proportion of click like
> p_pool <- (number_like_tips + number_like_tools)/(visitors_tips + visitors_tools)
> SE_pool<- sqrt(p_pool*(1-p_pool) * ((1/visitors_tips) + (1/visitors_tools)))
> d_hat <- phat_like_tips - phat_like_tools #Point Estimate or Difference in proportion
> z_score <- d_hat/SE_pool
> p_value <- pnorm(q = -z_score, mean = 0, sd = 1) * 2
> #Run a 2-sampled test
> print("H0:  proportion of click like(tips) = proportion of click like(tools)")
[1] "H0:  proportion of click like(tips) = proportion of click like(tools)"
> print("H1:  proportion of click like(tips) > proportion of click like(tools)")
[1] "H1:  proportion of click like(tips) > proportion of click like(tools)"
> prop.test(c(number_like_tips, number_like_tools), c(visitors_tips,visitors_tools))

	2-sample test for equality of proportions with continuity correction

data:  c(number_like_tips, number_like_tools) out of c(visitors_tips, visitors_tools)
X-squared = 681.57, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.08992453 0.10447547
sample estimates:
    prop 1     prop 2 
0.16626667 0.06906667 

> print("result p-value < 2.2e-16")
[1] "result p-value < 2.2e-16"
> 
```

>>作法：接著我們計算2種不同condition的按讚率，經計算發現使用tips這種condition此按讚率高於tools法140%。因此建立一個假設檢定，來檢定使用tips的寫作方式，是否在統計上能夠顯著足夠證明tips的高於差距是有意義的。而在2-sample test中，從檢定結果我們可以看到p-value < 0.05的結果，我們可以宣稱在alpha=0.05的level我們拒絕虛無假設。在統計上的意義也顯示其粉絲在**使用tips方式寫作的文章，按讚率有顯著高於使用tools寫作方式**。

>>分析建議：因而我們建議此網紅，想要提升文章按讚率，可以多嘗試使用tips的方式進行寫作。

附錄: R 程式碼

```
#HW6

setwd("~/Downloads/1102 R/HW/hw 6")
library(readr)
library(tidyverse)
data <- read.csv("hw6-fb.csv")
summary(data)

##變數類別整理
data <- data[,-1]
data$visit_date <- as.Date(data$visit_date)
data$condition <- as.factor(data$condition)
data$clicked_article <- as.factor(data$clicked_article)
data$clicked_like <- as.factor(data$clicked_like)
data$clicked_share <- as.factor(data$clicked_share)
data$gender <- as.factor(data$gender)

summary(data)

## condition difference
data %>%
  group_by(condition) %>%
  summarise(time_spent = mean(time_spent_homepage_sec))

## gender difference
data %>%
  group_by(gender) %>%
  summarise(time_spent = mean(time_spent_homepage_sec))


## （文章觀看停留時間分析）Hypothesis Test:
### t-test for two sample mean
### Ha:  mu_1(tips) - mu_(tools) >0
t.test(data[data$condition == "tips", ]$time_spent_homepage_sec,
       data[data$condition == "tools", ]$time_spent_homepage_sec,
       alternative = "greater")
#### conclude H0, the p-value does not less than the significance
at the level of 0.05
#### tools 與 tips 的寫作方式並沒有顯著差距影響讀者在頁面停留時間

##使用ggplot 畫圖展示
ggplot(data, aes(x = condition, y = time_spent_homepage_sec)) +
  geom_boxplot() +
  xlab("Condition") + ylab("time_spent_homepage_sec") +
  ggtitle("Boxplot of time_spent_homepage_sec by Condition") +
  theme_bw()
```

```
##2種condition對於文章按讚率分析
# condition-tips: proportion of like
like_tips <- data %>% filter(condition == "tips" & clicked_like ==
"1")
number_like_tips <- nrow(like_tips)
visitors_tips <- nrow(data %>% filter(condition == "tips"))
phat_like_tips <-  (number_like_tips/visitors_tips)

# condition-tools: proportion of like
like_tools <- data %>% filter(condition == "tools" & clicked_like
== "1")
number_like_tools <- nrow(like_tools)
visitors_tools <- nrow(data %>% filter(condition == "tools"))
phat_like_tools <-  (number_like_tools/visitors_tools)

##計算tips的按讚率高於tools的多少
uplift <- (phat_like_tips - phat_like_tools)/ phat_like_tools *
100
uplift  #140.74%
#tips的按讚率 is better than tools by 140%.

#pooled proportion of click like
p_pool <- (number_like_tips + number_like_tools)/(visitors_tips +
visitors_tools)
SE_pool<- sqrt(p_pool*(1-p_pool) * ((1/visitors_tips) + (1/
visitors_tools)))
d_hat <- phat_like_tips - phat_like_tools #Point Estimate or
Difference in proportion
z_score <- d_hat/SE_pool
p_value <- pnorm(q = -z_score, mean = 0, sd = 1) * 2


#Run a 2-sampled test
print("H0:  proportion of click like(tips) = proportion of click
like(tools)")
print("H1:  proportion of click like(tips) > proportion of click
like(tools)")

prop.test(c(number_like_tips, number_like_tools),
c(visitors_tips,visitors_tools))
print("result p-value < 2.2e-16")
```