統計學一下 期末專案

影響台大學生抽獎意願之因素

第一組

B07705008 巫芊瑩 B08705005 杜沛慈 B08705007 林又昕 B08705018 莊莊
B08705053 葉小漓 B08705026 陳沛妤 B08705031 陳沛竹 B08705037 翁子婷

一、動機與背景

台大校園內充滿了由各系所以及學生社團舉辦的活動，例如：週、營隊、之夜……等等，為了吸引潛在的參與者或觀眾，主辦方通常會與商家合作舉行抽獎，以增強宣傳的效果。另外，不少台大學生在需要蒐集大量資料或群體意見時，也會設計問卷於網路上徵求填答，並以抽獎的方式作為誘因，進而提升填答率。

我們在臉書的「NTU 台大學生交流板」上，觀察到與抽獎相關的貼文，發文者大多會要求抽獎者於貼文下方留言作為參加證明，並於留言區抽出得獎者。然而，每一篇貼文的留言數量卻不盡相同，少則僅有個位數的留言，多則達到兩、三百則的留言。這也間接指出抽獎活動為發文者目的（活動宣傳、徵求問卷填答）所帶來的效益，其實有著大範圍的變異。

為最大化抽獎活動的效益，我們想探討哪些因素會影響台大學生的抽獎意願，因此將主題訂為「影響台大學生抽獎意願的因素」，並建立複迴歸模型，以及進行 Kruskal-Wallis Test、Wilcoxon Rank Sum Test 的檢定，希望分析結果可以作為未來舉辦抽獎活動時的參考依據。

二、資料蒐集

我們將資料蒐集分成兩個部分，一是人工蒐集 NTU 台大學生交流板上自 2019 年至 2021 年 5 月，與抽獎相關貼文的資料。蒐集的內容包括：發文日期、留言數量、抽獎條件（是否需要填寫問卷，是記為 1，否記為 2）、獎品總價格、留言需標註的人數，以及貼文按讚數。資料筆數總共 214 筆，範例資料如下表所示。

date	comments	condition	total_price	tag	likes
2021/05/31	155	1	700	3	63
2021/05/31	48	1	200	2	43
2021/05/31	46	1	800	2	106

第二個部分是問卷蒐集，我們設計了一份問卷並將其發佈於 NTU 台大學生交流板，詢問台大學生對於各抽獎活動不同的內容，其抽獎意願為何（以 1～10 分計，越高分表示抽獎意願越高），問卷開放時間為 8 天。每份問卷填答的資料包括時間戳記、填答者的性別，以及針對抽獎內容的三個面向（獎品類型、獎品數量、獎品價格）分類，不同類別所對應到的抽獎意願。詳細的分類項目如以下各表所示，資料筆數共有 105 筆。

1. 時間戳記與填答者性別

時間戳記	性別
5/31/2021 20:12:19	生理男
5/31/2021 20:16:30	生理男
5/31/2021 20:21:02	生理男

2. 獎品類型

食品飲料	優惠券\禮品卡\票券	文創品	衣服	書籍	彩妝	電子產品	抱枕、娃娃、療癒小物
8	6	6	8	5	1	8	8
7	5	6	4	6	1	10	4
6	1	5	6	6	6	7	1

3. 獎品數量

1 to 3	3 to 5	more than 5
9	9	10
5	6	6
3	5	7

4. 獎品價格

less than 100	100 to 400	400 to 600	600 to 800	more than 800
4	6	8	9	10
8	9	10	10	10
1	3	6	7	8

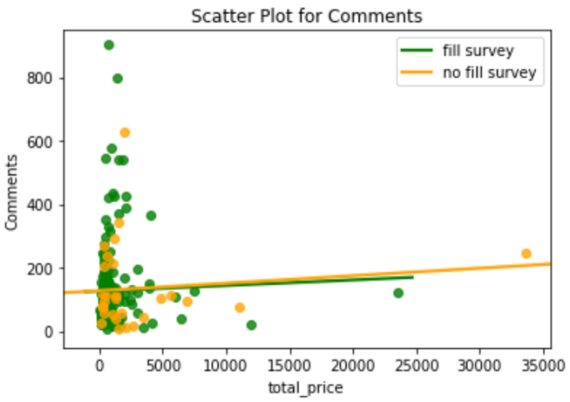
三、資料描述

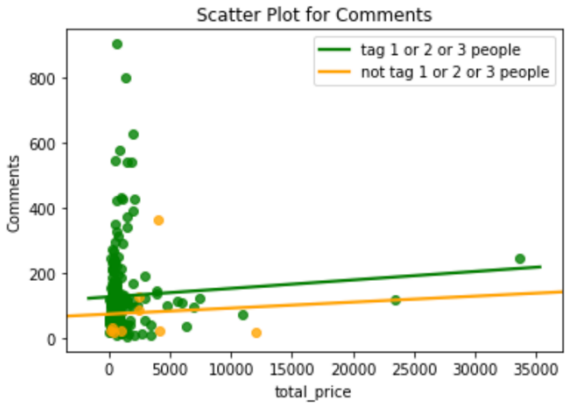
1. 人工蒐集資料

	comments	condition	total_price	tag	likes
count	214.000000	214.000000	214.000000	214.000000	214.000000
mean	129.177570	1.135514	1293.355140	2.182243	82.616822
std	131.655748	0.343074	3126.725592	0.756482	68.113154
min	7.000000	1.000000	70.000000	0.000000	13.000000
25%	50.250000	1.000000	300.000000	2.000000	41.000000
50%	94.500000	1.000000	500.000000	2.000000	62.000000
75%	145.250000	1.000000	1000.000000	3.000000	100.250000
max	903.000000	2.000000	33602.000000	5.000000	469.000000

由上表可知，第一欄顯示抽獎貼文的留言數量範圍是 7~903 則，平均留言數量為 129 則；第二欄的第三四分位數為 1，表示至少有 75%的抽獎貼文需要填寫問卷；第三欄顯示抽獎貼文的獎品總價格範圍是 70~33602 元，平均獎品總價格為 1293 元；第四欄顯示抽獎貼文的標註人數範圍是 0~5 人，平均標註人數為 2 人；第五欄顯示抽獎貼文的按讚數量範圍是 13~469 個，平均按讚數量為 83 個。

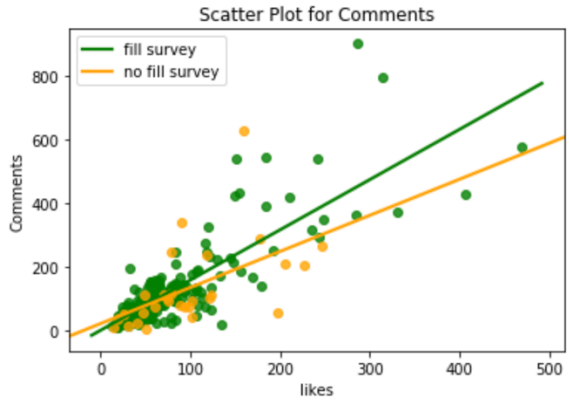
下方兩張圖為獎品總價格對留言數量的散佈圖，資料點的分類方式有兩種，左邊的圖是依據抽獎條件分類，右邊的圖則是依據標註人數分類，在兩張圖中獎品總價格與留言數量的線性關係非常不明顯。

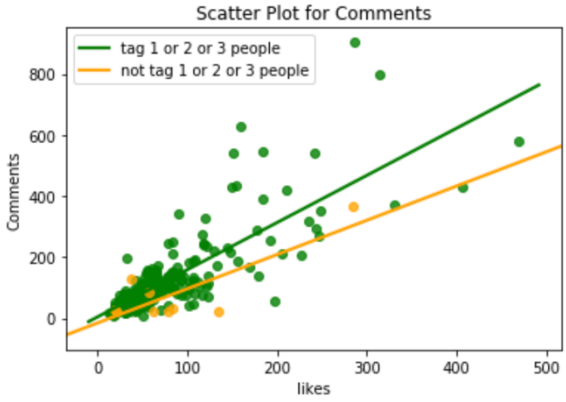




由左圖可以看到，是否需要填問卷的抽獎活動，其留言數量及獎品總價格的分布狀況並沒有太大分別。右圖則顯示出，需標註 1~3 人的抽獎貼文，其留言數量高於不需標註他人或需標註 3 人以上的貼文。

下方兩張圖為按讚數量對留言數量的散佈圖，資料點的分類方式有兩種，左邊的圖是依據抽獎條件分類，右邊的圖則是依據標註人數分類，我們可以觀察到按讚數量與留言數量有比較明顯的線性關係。



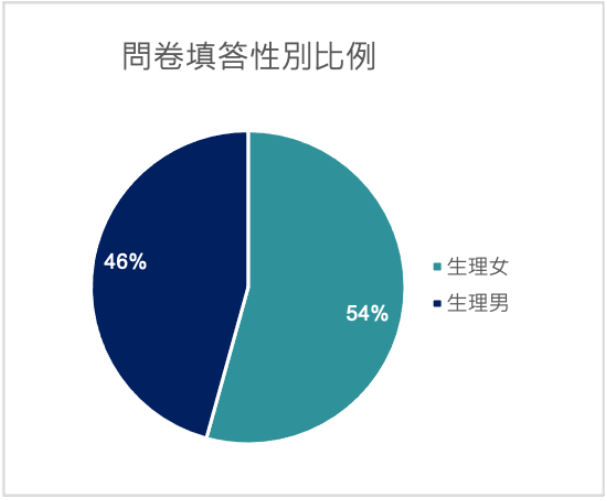


在左圖中，是否需要填問卷的抽獎活動，其留言數量和按讚數量的分布狀況並沒有太大分別，兩條迴歸線交叉的情況表示抽獎條件與按讚數量間可能存在交互作用。右圖中不需標註他人或需標註 3 人以上的貼文，其留言數量與按讚數量比需要標註 1~3 人的貼文，皆相對較少。

2. 問卷資料

(1) 填答者性別

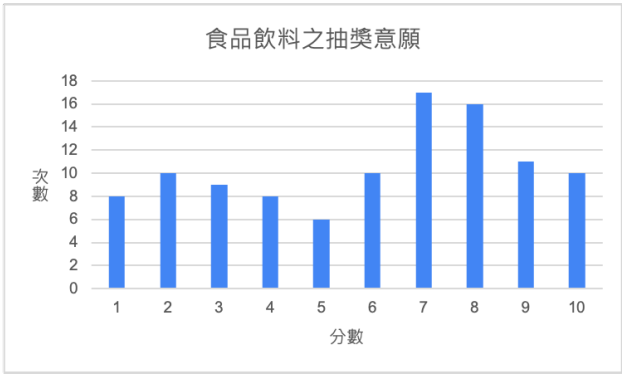
本問卷共有 57 位女性、48 位男性填答，性別比例如下圖所示。

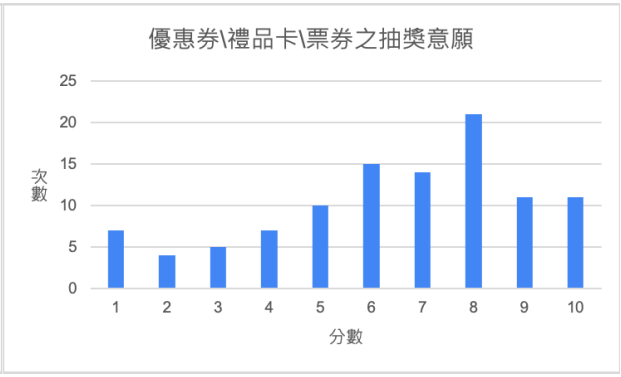


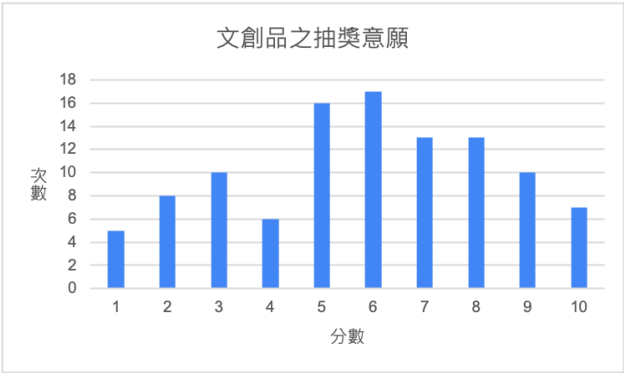
(2) 獎品類型

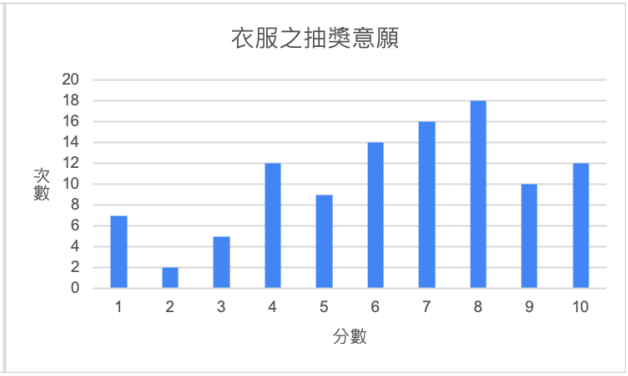
	食品飲料	優惠券\禮品卡\票券	文創品	衣服	書籍	彩妝	電子產品	抱枕、娃娃、療癒小物
count	105.000000	105.000000	105.000000	105.000000	105.000000	105.000000	105.000000	105.000000
mean	5.933333	6.409524	5.828571	6.371429	5.990476	4.609524	8.152381	5.752381
std	2.805443	2.552130	2.494059	2.527758	2.486484	3.089981	1.945315	2.820873
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	3.000000	5.000000	4.000000	5.000000	4.000000	1.000000	7.000000	4.000000
50%	7.000000	7.000000	6.000000	7.000000	6.000000	4.000000	8.000000	6.000000
75%	8.000000	8.000000	8.000000	8.000000	8.000000	7.000000	10.000000	8.000000
max	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000

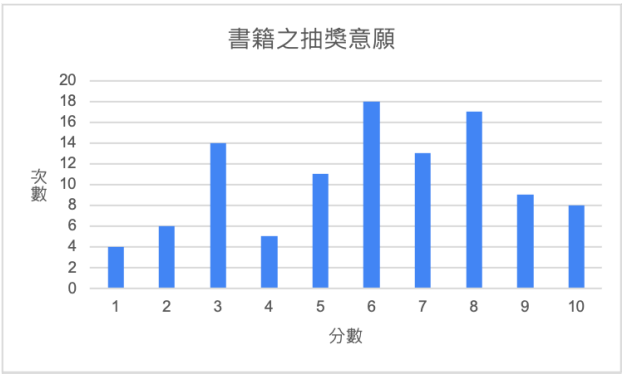
我們可以由上表看到，在八類獎品中，抽獎意願最高的是電子產品，平均抽獎意願為 8.15；抽獎意願最低的是彩妝品，平均抽獎意願僅有 4.61。以標準差來看，變異最小的是電子產品，標準差為 1.95；變異最大的為彩妝品，標準差為 3.09。以下八張長條圖為各獎品種類之抽獎意願分佈。

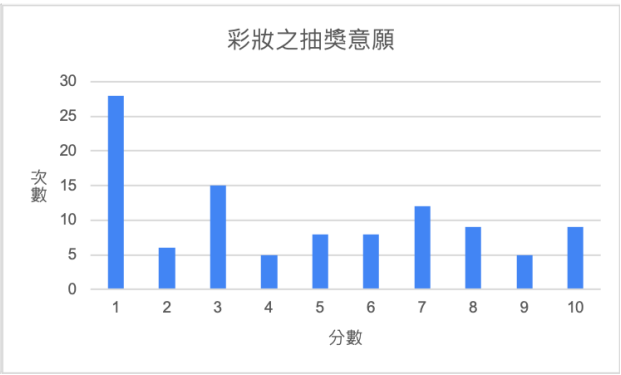


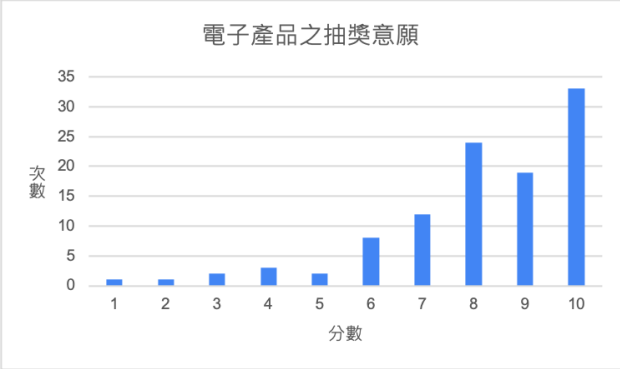


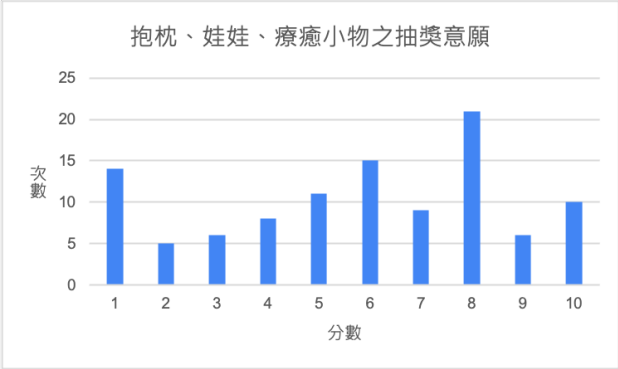








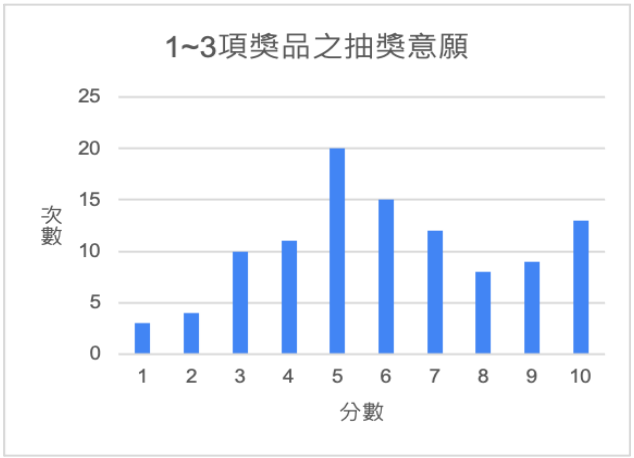


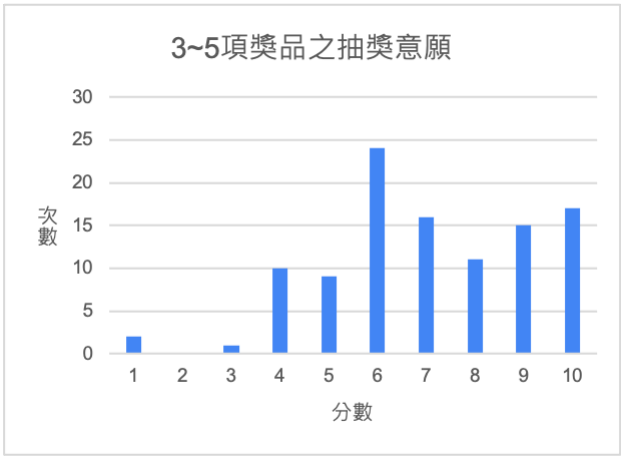


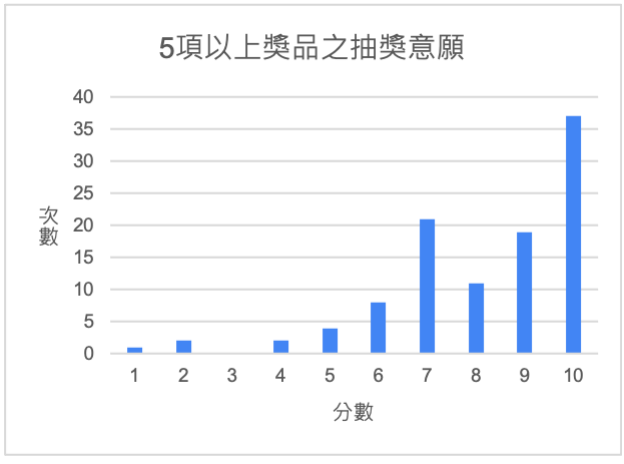
(3) 獎品數量

	number							
	count	mean	std	min	25%	50%	75%	max
factor								
1to3	105.0	6.038095	2.453113	1.0	4.0	6.0	8.0	10.0
3to5	105.0	7.038095	2.107331	1.0	6.0	7.0	9.0	10.0
5more	105.0	8.161905	1.976418	1.0	7.0	9.0	10.0	10.0

在獎品數量方面，1~3 項獎品的平均抽獎意願為 6.04，3~5 項獎品的平均抽獎意願為 7.04，5 項以上獎品的平均抽獎意願為 8.16，抽獎意願隨獎品數量呈現上升的趨勢。以下三張長條圖為各獎品數量其抽獎意願的分佈。



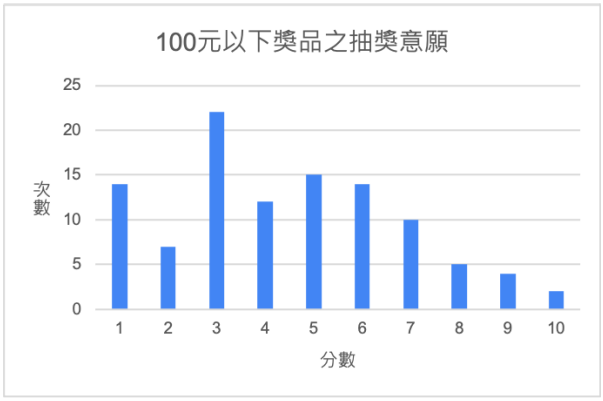


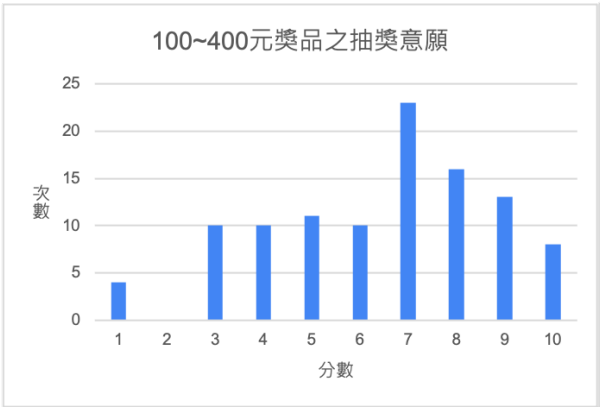


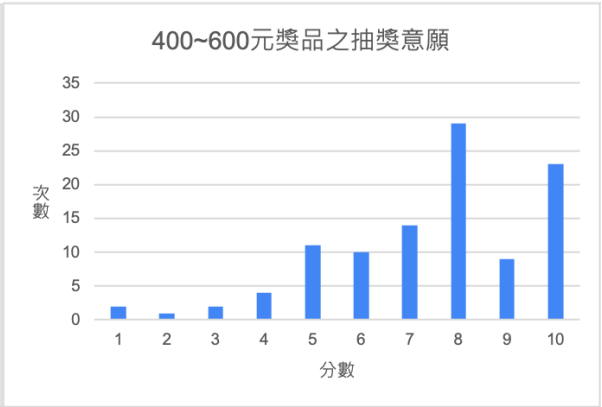
(4) 獎品價格

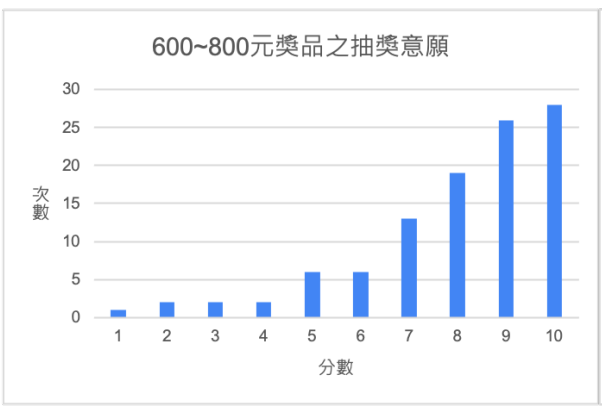
	price							
	count	mean	std	min	25%	50%	75%	max
factor								
100less	105.0	4.447619	2.336915	1.0	3.0	4.0	6.0	10.0
100to400	105.0	6.428571	2.307418	1.0	5.0	7.0	8.0	10.0
400to600	105.0	7.447619	2.125762	1.0	6.0	8.0	9.0	10.0
600to800	105.0	8.019048	2.038009	1.0	7.0	9.0	10.0	10.0
800more	105.0	8.828571	1.913802	1.0	8.0	10.0	10.0	10.0

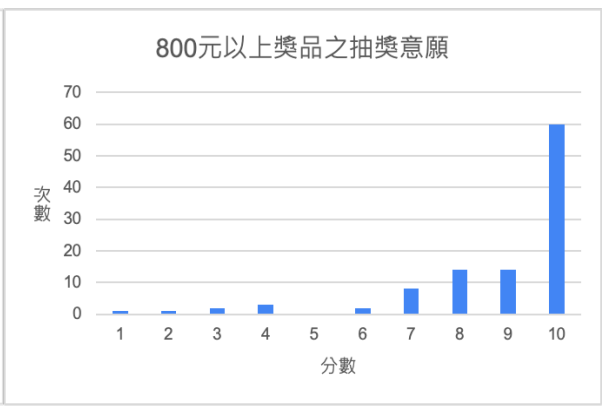
在獎品價格方面，100 元以下獎品的平均抽獎意願為 4.45，100~400 元獎品的平均抽獎意願為 6.43，400~600 元獎品的平均抽獎意願為 7.45，600~800 元獎品的平均抽獎意願為 8.02，800 元以上獎品的平均抽獎意願為 8.83，抽獎意願隨獎品數量呈現上升的趨勢。以下五張長條圖為各獎品價格其抽獎意願的分佈。











四、人工蒐集資料分析－複迴歸模型

由於我們想要了解各個因素之間與抽獎貼文留言數量是否相關，以及其相關的方向和強度，因此我們將人工蒐集的資料用於建立一個複迴歸模型。人工蒐集資料分析的詳細過程皆記錄於檔案「Stat_2020_b_1_Regression.ipynb」中，模型的變數定義與迴歸線方程式如下：

y：留言數量

x_1：獎品總價格

x_2：按讚數量

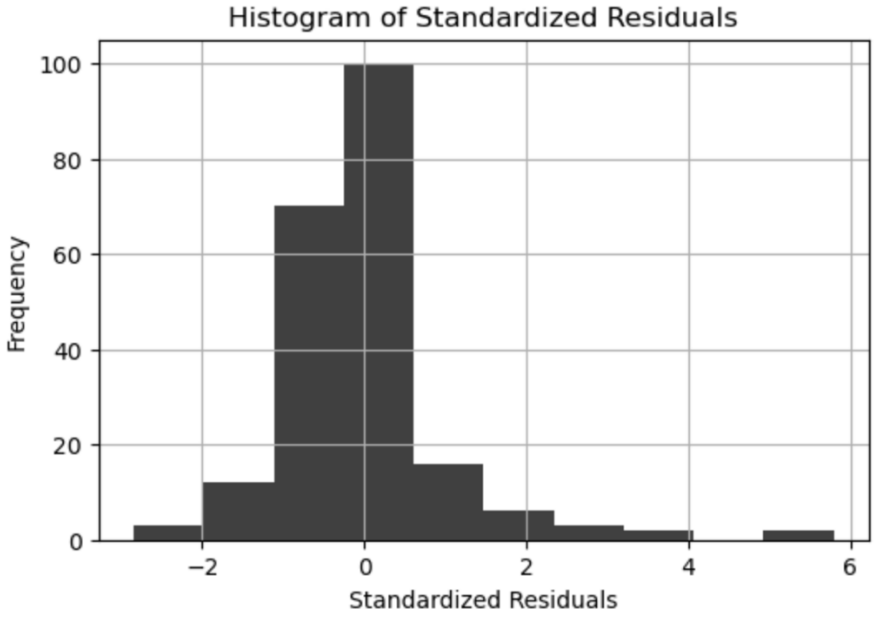
x_3：抽獎條件（為類別變數，需填寫問卷記為 1，否則為 0)

x_4：標註人數（為類別變數，標註 1~3 人記為 1，其餘為 0)

迴歸線：$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 + \varepsilon$

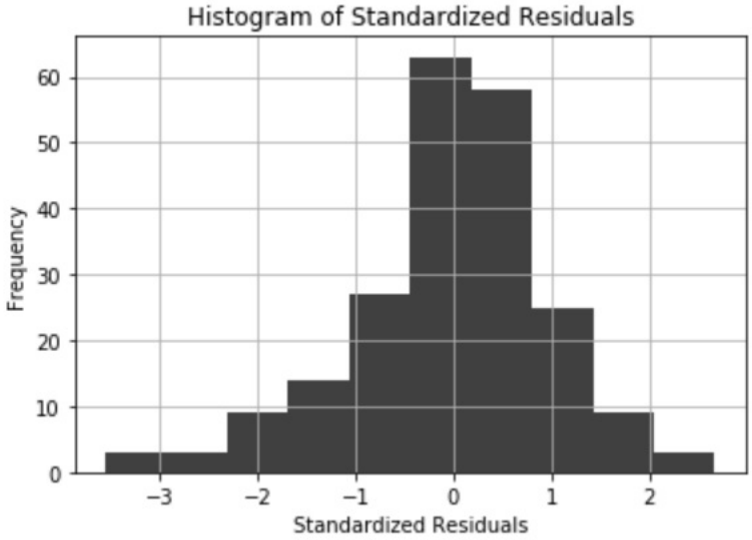
其中β_i為未知且需估計的迴歸係數，ε為誤差項。

我們在電腦跑出第一次模型後，發現殘差的直方圖分佈呈現右斜的狀態，且沒有辦法滿足常態的假設（Shapiro 檢定的顯著性 p 值趨近於 0），下圖為殘差的直方圖以及 Shapiro 檢定的結果。



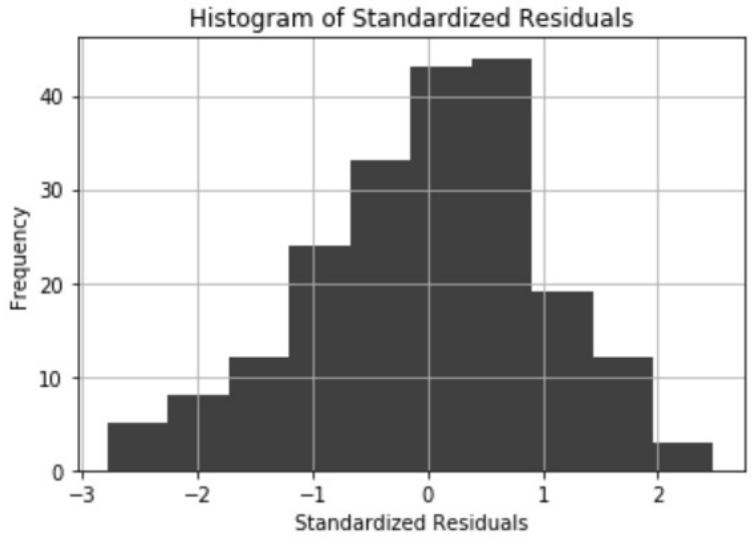

```
[  3.  12.  70. 100.  16.   6.   3.   2.   0.   2.]
[-2.8424348  -1.97861451 -1.11479422 -0.25097393  0.61284635  1.47666664
  2.34048693  3.20430722  4.0681275   4.93194779  5.79576808]
Shapiro Test: Statistics=0.812380, p=0.000000
```

因此我們將留言數量進行 y’ = log y 的變數轉換，變數變換後模型的殘差直方圖及 Shapiro 檢定結果如下圖所示，仍未滿足常態假設。




```
[ 3.  3.  9. 14. 27. 63. 58. 25.  9.  3.]
[-3.55252114 -2.93035822 -2.30819531 -1.6860324  -1.06386949 -0.44170657
  0.18045634  0.80261925  1.42478216  2.04694507  2.66910799]
Shapiro Test: Statistics=0.972861, p=0.000383
```

先刪除離群值，再進行一次分析。殘差直方圖及 Shapiro 檢定結果如下圖所示，可滿足常態假設。




```
[ 5.  8. 12. 24. 33. 43. 44. 19. 12.  3.]
[-2.77822798 -2.2521175  -1.72600701 -1.19989653 -0.67378605 -0.14767557
  0.37843491  0.90454539  1.43065587  1.95676635  2.48287683]
Shapiro Test: Statistics=0.987325, p=0.067216
Chi-squared test: statistics = 0.9606, p-value = 0.3270
Critical value = 3.8415 (defree of freedom = 1)
```

進行其他必要條件的檢定，發現樣本在 Run Test 中不滿足隨機性的假設，因此加入時間序列作為自變數以解決此問題，第四次分析的模型及結果如下表所示。

y：留言數量

x_1：獎品總價格

x_2：按讚數量

x_3：抽獎條件（為類別變數，需填寫問卷記為 1，否則為 0)

x_4：標註人數（為類別變數，標註 1~3 人記為 1，其餘為 0)

x_5：時間序列

迴歸線：$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$

其中β_i為未知且需估計的迴歸係數，ε為誤差項。

OLS Regression Results

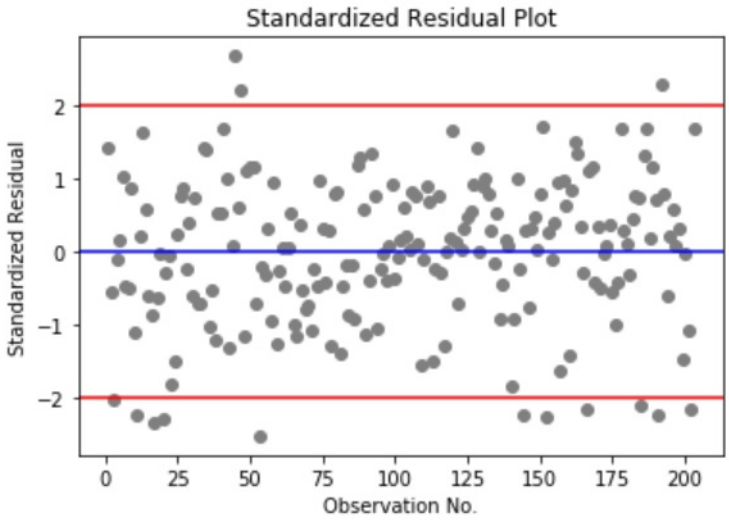
Dep. Variable:	comments_log	R-squared:	0.640
Model:	OLS	Adj. R-squared:	0.630
Method:	Least Squares	F-statistic:	69.92
Date:	Tue, 22 Jun 2021	Prob (F-statistic):	8.53e-42
Time:	15:11:52	Log-Likelihood:	-136.96
No. Observations:	203	AIC:	285.9
Df Residuals:	197	BIC:	305.8
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.2784	0.247	9.213	0.000	1.791	2.766
total_price	2.726e-05	1.14e-05	2.385	0.018	4.72e-06	4.98e-05
likes	0.0102	0.001	17.405	0.000	0.009	0.011
condition_1	0.3528	0.109	3.250	0.001	0.139	0.567
tags_123	0.9545	0.189	5.062	0.000	0.583	1.326
Week_Number	0.0049	0.002	2.251	0.026	0.001	0.009

Omnibus:	3.023	Durbin-Watson:	1.902
Prob(Omnibus):	0.221	Jarque-Bera (JB):	3.020
Skew:	-0.294	Prob(JB):	0.221
Kurtosis:	2.893	Cond. No.	2.98e+04

預估模型：$\hat{y} = 2.2784 + 2.726e^{-05} x_1 + 0.0102 x_2 + 0.3528 x_3 + 0.9545 x_4 + 0.0049 x_5$

此模型的 R 平方值為 0.640，調整後的 R 平方值為 0.630，表示迴歸線與資料相符。進行隨機性檢定，不拒絕 Run test 的虛無假設，顯示樣本為隨機選擇，如下圖所示。




```
runs =  93
n1 =  102
n2 =  101
runs_exp =  102.49753694581281
stan_dev =  7.1060748203268504
z =  -1.3365377069553517
pval_z =  0.18137358434578132
p_value for Z-statistic=  0.18137358434578132
```

在評估模型可信度的階段，我們以 t 檢定來檢定係數，可知所有自變數與留言數量皆有線性關係。

模型的係數指出，當獎品總價格上升 1 元時，留言數量會上升{2.726e}^{-05} 筆；當按讚數量增加 1 時，留言數量會上升 0.0102 筆；需填寫問卷的抽獎活動，其留言數量會大於不需填寫問卷的抽獎活動；需標註 1 至 3 人的抽獎貼文，其留言數量會大於不需標註其他人或需標註更多人的抽獎貼文。

我們由複迴歸模型的結果得知，按讚數量、標註人數為與留言數量間有較強關係的前二項因素。因此我們可以建議抽獎主辦方，當他們想要提高抽獎活動的效益時，可依此考量並設定抽獎條件，進而吸引更多抽獎者。

五、問卷資料分析－Kruskal-Wallis Test、Wilcoxon Rank Sum Test

我們將問卷蒐集到的資料用於三個 Kruskal-Wallis 檢定，以及一個 Wilcoxon Rank Sum Test 的檢定。問卷資料分析的詳細過程皆記錄於檔案「Stat_2020_b_1_ANOVA.ipynb」中。

1. 獎品類型、抽獎者性別

由於我們原先想要了解不同性別的抽獎者對於獎品類型是否會有不同的偏好，因此決定進行 Two-Way ANOVA 檢定來確認「獎品種類」與「抽獎者性別」之間是否存在交互作用，進而導致較高或較低的抽獎意願。檢定的假設如下：

H_0：獎品種類與抽獎者性別之間不存在影響抽獎意願的交互作用
H_1：獎品種類與抽獎者性別之間存在影響抽獎意願的交互作用

	sum_sq	df	F	PR(>F)
C(form)	717.138095	7.0	15.957302	9.672708e-20
C(sex)	147.392798	1.0	22.957837	1.963804e-06
C(form):C(sex)	226.859301	7.0	5.047929	1.271536e-05
Residual	5290.205044	824.0	NaN	NaN

檢定結果如上表所示，前兩列分別代表獎品種類（form）及抽獎者性別（sex），第三列代表的是兩因子的交互作用項，其 F 統計值為 5.0479，顯著性 p 值為 0.00001，小於 0.05。因此我們拒絕了虛無假設，可以說明獎品種類與抽獎者性別之間存在交互作用，特定兩因子的組合可能導致較高或較低的抽獎意願。

然而我們在檢驗 Two-Way ANOVA 的必要條件時，由於母群體為常態分佈、變異數同質性的假設都沒有被滿足，該檢定的結果可能不具備參考價值。因此我們決定使用 Kruskal-Wallis Test 及 Wilcoxon Rank Sum Test 來個別檢驗，「獎品種類」和「抽獎者性別」是否為影響抽獎意願的主要因素。

(1) 獎品種類－Kruskal-Wallis Test

令八種獎品分別為八個母體，檢定的假設如下：

H_0：八個母體的位置皆相同

H_1：至少有兩個母體的位置不相同

```
H = 93.74236250980721
p-value = 0.0
```

檢定結果如上所示，H 值為 93.74，顯著性 p 值趨近於 0，因此我們可以拒絕虛無假設，並得知至少有兩個母體的位置不相同。整體而言，依據獎品種類的不同，參加抽獎的意願會有差別。

(2) 抽獎者性別－Wilcoxon Rank Sum Test

H_0：男性的抽獎意願不比女性的抽獎意願低

H_1：男性的抽獎意願比女性的抽獎意願低

```
MannwhitneyuResult(statistic=73652.5, pvalue=3.245149836011175e-05)
```

檢定結果如上所示，顯著性 p 值趨為 0.00003，因此我們可以拒絕虛無假設，並得知男性的抽獎意願比女性的抽獎意願低。

2. 獎品數量、獎品價格

(1) 獎品數量

我們想要了解抽獎活動的獎品數量是否為影響抽獎意願之主要因素之一，因此我們將獎品數量分為三組（1~3 項、3~5 項、5 項以上），並進行了 One-Way ANOVA 的檢定，檢定的假設如下：

H_0：各組獎品數量的平均抽獎意願皆相等

H_1：至少有兩組之間的平均抽獎意願不相等

	sum_sq	df	F	PR(>F)
C(factor)	237.073016	2.0	24.755559	1.049876e-10
Residual	1493.942857	312.0	NaN	NaN

檢定結果如上表所示，第一列代表的是獎品數量，其 F 統計值為 24.7556，顯著性 p 值趨近於 0。因此我們可以拒絕虛無假設，並推知抽獎活動的獎品數量的確會影響抽獎意願。

然而我們在檢驗 One-Way ANOVA 的必要條件時，由於母群體為常態分佈的假設沒有被滿足，該檢定的結果可能不具備參考價值。因此我們決定使用 Kruskal-Wallis Test 來檢驗獎品數量是否為影響抽獎意願的主要因素。令獎品數量的區間（1~3 項、3~5 項、5 項以上）分別為三個母體，檢定的假設如下：

H_0：三個母體的位置皆相同
H_1：至少有兩個母體的位置不相同

```
H = 44.009479606188506
p-value = 2.776278096305873e-10
```

檢定結果如上所示，H 值為 44.01，顯著性 p 值趨近於 0，因此我們可以拒絕虛無假設，並得知至少有兩個母體的位置不相同。整體而言，依據獎品數量的不同，參加抽獎的意願會有差別。

(2) 獎品價格

由於我們想要了解抽獎活動的獎品價格是否對於抽獎意願有顯著的影響，因此我們將獎品價格分為五組（100 元以下、100~400 元、400~600 元、600~800 元、800 元以上），進行 One-Way ANOVA 的檢定，檢定的假設如下：

H_0：各組獎品價格的平均抽獎意願皆相等
H_1：至少有兩組之間的平均抽獎意願不相等

	sum_sq	df	F	PR(>F)
C(factor)	1198.868571	4.0	64.816797	1.838375e-44
Residual	2404.514286	520.0	NaN	NaN

檢定結果如上表所示，第一列代表的是獎品價格，其 F 統計值為 64.8168，顯著性 p 值趨近於 0。因此我們可以拒絕虛無假設，並推得抽獎活動的獎品價格是影響抽獎意願的主要因素之一。

然而我們在檢驗 One-Way ANOVA 的必要條件時，由於母群體為常態分佈的假設沒有被滿足，該檢定的結果可能不具備參考價值。因此我們決定使用 Kruskal-Wallis Test 來檢驗獎品價格是否為影響抽獎意願的主要因素。令獎品數量的區間（100 元以下、100~400 元、400~600 元、600~800 元、800 元以上）分別為五個母體，檢定的假設如下：

H_0：五個母體的位置皆相同
H_1：至少有兩個母體的位置不相同

```
H = 175.02451421328988
p-value = 0.0
```

檢定結果如上所示，H 值為 175.02，顯著性 p 值趨近於 0，因此我們可以拒絕虛無假設，並得知至少有兩個母體的位置不相同。整體而言，依據獎品價格的不同，參加抽獎的意願會有差別。

六、結論與建議

我們本次研究的主要對象限定為台大學生，對於一些學生舉辦的抽獎活動，可能沒有辦法提供數量龐大或價值很高的獎品，而我們藉由這次的研究，發現在學生負擔得起的價位範圍內，即使只是小幅度地增加獎品的價位和數量，也能讓參與者有不同的抽獎意願，因此，我們建議抽獎主辦方在預算內盡量提升獎品的數量與價值，理應能為吸引抽獎者帶來很大的助益。

另外，我們發現依據獎品的不同，學生參加抽獎的意願度會有所差別，其中又以電子產品作為獎品時的意願度最高。並且，整體而言，男學生參加抽獎的意願度相較女學生來得低，因此我們也建議若活動主辦方的目標對象以女性為主時，較適合用抽獎方式吸引參與者，而當目標對象是男性時，就可以考慮採用其他宣傳方案。

而為了讓台大學生們在「NTU 台大學生交流板」，能有效地達到舉辦抽獎活動的目的，如：活動宣傳、徵求問卷填答等等，我們藉由回歸分析探討影響台大學生在抽獎文底下留言的因素，發現我們所選定的因素中，按讚數量、標註人數兩項與留言數量之間有較強的正向線性關係。因此我們建議抽獎主辦方可以藉由要求參與者按讚並標註朋友，進而吸引更多人在抽獎貼文留言。