大數據分析與實作_期末作業

指導教授：龔千芬 老師

組別：第十一組

組員：

C108156122 王薈宣

C108156135 鄭繼威

C108156140 葉桔良

C108156144 吳岳峰

第一題

1. 某海產業者想知道雌雄鮑魚重量有無差異？

2. Data Source：https://archive.ics.uci.edu/ml/datasets/Abalone

Gender	Length	Diameter	Height	Whole wei	Shucked w	Viscera wi	Shell weig	Rings
M	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15
M	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7
F	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9
M	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10
I	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7
I	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8
F	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20
F	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16
M	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9
F	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19
F	0.525	0.38	0.14	0.6065	0.194	0.1475	0.21	14
M	0.43	0.35	0.11	0.406	0.1675	0.081	0.135	10
M	0.49	0.38	0.135	0.5415	0.2175	0.095	0.19	11
F	0.535	0.405	0.145	0.6845	0.2725	0.171	0.205	10
F	0.47	0.355	0.1	0.4755	0.1675	0.0805	0.185	10
M	0.5	0.4	0.13	0.6645	0.258	0.133	0.24	12
I	0.355	0.28	0.085	0.2905	0.095	0.0395	0.115	7
F	0.44	0.34	0.1	0.451	0.188	0.087	0.13	10
M	0.365	0.295	0.08	0.2555	0.097	0.043	0.1	7
M	0.45	0.32	0.1	0.381	0.1705	0.075	0.115	9

資料描述：性別分為 F（公）、M（母）、I（嬰兒無法分辨）
重量指的是一顆鮑魚帶殼的重量單位為公斤

3. 分析方法：獨立樣本 T 檢定

為何採用：X 為性別（名目尺度）、Y 為鮑魚含殼重量（連續尺度）

4. 分析結果：

程式碼（Abalone_ind.py）：

```
from scipy import stats
import pandas as pd

df1 = pd.read_excel('Abalone.xlsx')
# print(df1)
F = []
M = []
I = []
```

```
for i in range(len(df1['Gender'])):
    if df1['Gender'][i] == "F":
        F.append(df1['Whole_weight'][i])
    elif df1['Gender'][i] == "M":
        M.append(df1['Whole_weight'][i])
    else:
        I.append(df1['Whole_weight'][i])

lev = stats.levene(F, M, center='mean')
print('lev : ', lev)
if lev[1] > 0.05:
    ay6t = stats.ttest_ind(F, M)
    print('變異數同質 : ', ay6t)
else:
    ay6tf = stats.ttest_ind(F, M, equal_var=False)
    print('變異數異質 : ', ay6tf)
```

統計假設：

H0：μ雌 = μ雄

H1：μ雌 != μ雄

5. 結論：

```
lev :  LeveneResult(statistic=5.2153910347064825, pvalue=0.022461508564473653)
變異數異質 :  Ttest_indResult(statistic=3.2530891638844825, pvalue=0.0011550413625208316)
```

P = 0.0011550413625208316 < α = 0.05

拒絕 H0 =>

在α = 0.05，雌雄鮑魚重量無差異的說法是不可以支持的。

第二題

1. 某飲料業者想知道夏季與冬季時在飲料銷售（杯數）有無差異？

2. Data Source：https://reurl.cc/bkkGRr

winterdate	wintercup	winterlowtemp	summerdate	summercup	summerlowtemp
11/22	98	20	9/18	277	27
11/23	103	19	9/19	243	28
11/24	116	21	9/20	231	27
11/25	121	22	9/21	212	28
11/26	127	21	9/22	215	29
11/27	137	24	9/23	231	31
11/28	148	23	9/24	208	30
11/29	134	23	9/25	217	32
11/30	96	17	9/26	233	31

資料描述：這裡面的資料是分別取飲料店夏天與冬天某幾天晚上販賣的杯數。

3. 分析方法：獨立樣本 T 檢定

為何採用：X 為季節（名目尺度），Y 為飲料杯數（連續尺度）

4. 分析結果：
程式碼（drinks.py）：

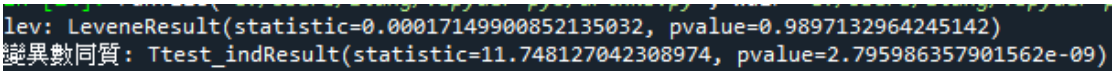
```
from scipy import stats
import pandas as pd

df1 = pd.read_excel('drinks.xlsx')
winterCup = df1['wintercup']
summerCup = df1['summercup']

lev = stats.levene(summerCup,winterCup,center='mean')
print('lev:',lev)

if lev[1] > 0.05:
    result = "變異數同質:"
    ttest = stats.ttest_ind(summerCup,winterCup)
else:
    result = "變異數異質:"
    ttest = stats.ttest_ind(summerCup,winterCup,equal_var=False)

print(result,ttest)
```

統計假設：

H0：μ冬季 = μ夏季

H1：μ冬季 != μ夏季

5. 結論：



P = 2.795986357901562e-09 < α = 0.05

拒絕 H0 =>

在α=0.05，兩季的飲料銷售數量無差異的說法是不可以支持的。

第三題

1. 在發生疫情前與發生疫情後各國 GDP 是否有差異？

2. Data Source：https://reurl.cc/Q6K7A2 & https://reurl.cc/82xyMM

排名	國家/地區	所在洲	GDP(美元)	佔世界%
	全世界		86.14萬億(86,139,293,063,112)	
1	美國	美洲	20.61萬億(20,611,860,934,000)	23.9285%
	歐盟地區		15.97萬億(15,970,405,599,565)	18.5402%
2	中國	亞洲	13.89萬億(13,894,817,549,380)	16.1306%
3	日本	亞洲	5.04萬億(5,036,891,740,656)	5.8474%
4	德國	歐洲	3.98萬億(3,975,347,237,442)	4.6150%
5	英國	歐洲	2.86萬億(2,857,316,524,862)	3.3171%
6	法國	歐洲	2.79萬億(2,789,593,979,064)	3.2385%
7	印度	亞洲	2.7萬億(2,701,111,782,774)	3.1357%
8	意大利	歐洲	2.09萬億(2,091,117,091,124)	2.4276%
9	巴西	美洲	1.92萬億(1,916,947,014,067)	2.2254%
10	韓國	亞洲	1.72萬億(1,724,845,615,629)	2.0024%
11	加拿大	美洲	1.72萬億(1,721,853,332,869)	1.9989%

排名	國家/地區	所在洲	GDP(美元)	佔世界%
	全世界		84.58萬億(84,577,962,952,008)	
1	美國	美洲	20.94萬億(20,936,600,000,000)	24.7542%
	歐盟地區		15.28萬億(15,276,468,991,757)	18.0620%
2	中國	亞洲	14.72萬億(14,722,730,697,890)	17.4073%
3	日本	亞洲	4.98萬億(4,975,415,241,562)	5.8826%
4	德國	歐洲	3.85萬億(3,846,413,928,653)	4.5478%
5	英國	歐洲	2.71萬億(2,707,743,777,173)	3.2015%
6	法國	歐洲	2.63萬億(2,630,317,731,455)	3.1099%
7	印度	亞洲	2.62萬億(2,622,983,732,006)	3.1013%
8	意大利	歐洲	1.89萬億(1,886,445,268,340)	2.2304%
9	加拿大	美洲	1.64萬億(1,644,037,286,481)	1.9438%
10	韓國	亞洲	1.63萬億(1,630,525,005,469)	1.9278%
11	俄羅斯	歐洲	1.48萬億(1,483,497,784,867)	1.7540%

資料描述：每個國家所對應疫情前後的 GDP

3. 分析方法：獨立樣本 T 檢定

為何採用：X 為年份（名目尺度），Y 為各國 GDP（連續尺度）

4. 分析結果：
程式碼（GDP_ind.py）：

```
from scipy import stats
import pandas as pd

df1 = pd.read_excel('GDP.xlsx')
GDP2018 = df1['2018GDP(美元)']
GDP2020 = df1['2020GDP(美元)']
# print(before)

lev = stats.levene(GDP2018, GDP2020, center='mean')
print('lev：', lev)
if lev[1] > 0.05:
    ay6t = stats.ttest_ind(GDP2018, GDP2020)
    print('變異數同質：', ay6t)
else:
    ay6tf = stats.ttest_ind(GDP2018, GDP2020, equal_var=False)
```

```
print('變異數異質：', ay6tf)
```

統計假設：

H0：各國 GDP 無差異

H1：各國 GDP 有差異

5. 結論：

```
lev：  LeveneResult(statistic=0.0020546956338708794, pvalue=0.9638711221622318)
變異數同質：  Ttest_indResult(statistic=0.046055615664130196, pvalue=0.9632921882773815)
```

P = 0.9632921882773815 > α = 0.05

接受 H0 =>

在α = 0.05，疫情前後 GDP 無差異的說法是可以支持的。

第四題

1. 某遊戲為了平衡，在角色技能做了一些調整，請問在更新前與更新後角色勝率是否有差異？

2. Data Source：https://tw.op.gg/champion/statistics

	11.22勝率	11.23勝率
Viktor	49.59%	52.68%
Camille	52.42%	51.33%
Vayne	50.23%	51.09%
Lrelia	49.16%	49.35%
Tahm Kench	52.93%	52.43%
Teemo	49.29%	50.01%
Jaycy	45.71%	48.41%
Jax	50.22%	50.99%
Fiora	50.37%	49.97%
Tryndamere	51.08%	53.05%
Malphite	50.95%	50.84%
Yone	47.72%	49.95%
Poppy	53.90%	52.42%

資料描述：分為 11.22 及 11.23 版本的角色勝率

3. 分析方法：獨立樣本 T 檢定

為何採用：X 為版本（名目尺度），Y 為各英雄勝率（連續尺度）

4. 分析結果：
程式碼（Hero_ind.py）：

```
from scipy import stats
import pandas as pd

df1 = pd.read_excel('hero.xlsx')
win1122 = df1['11.22勝率']
win1123 = df1['11.23勝率']
# print(before)

lev = stats.levene(win1122, win1123, center='mean')
print('lev：', lev)
if lev[1] > 0.05:
```

```
    ay6t = stats.ttest_ind(win1122, win1123)
    print('變異數同質：', ay6t)
else:
    ay6tf = stats.ttest_ind(win1122, win1123, equal_var=False)
    print('變異數異質：', ay6tf)
```

統計假設：

H0：μ更新前 = μ更新後

H1：μ更新前 != μ更新後

5. 結論：

```
lev：  LeveneResult(statistic=0.08337579371630507, pvalue=0.7733272051159776)
變異數同質：  Ttest_indResult(statistic=-0.18178972215622455, pvalue=0.8560885111177758)
```

P = 0.8560885111177758 > α = 0.05

接受 H0 =>

在α = 0.05，更新前後角色勝率無差異的說法是可以支持的。

第五題

1. 某文化部長想知道 109 年高雄市各語言分布比例是否與台灣各語言分布有差異？

2. Data Source：https://www.thenewslens.com/article/157030

	台灣	高雄
國語	0.6637	0.555
台語	0.3173	0.432
客語	0.015	0.01
原住民語	0.002	0.001
其他	0.002	0.002

資料描述：高雄市跟台灣的各語言使用比例

3. 分析方法：卡方適合度檢定

為何採用：高雄市的分布為期望值，台灣的分布為觀察值

4. 分析結果：
程式碼（Language_TK_chi.py）：

```
from scipy.stats import chisquare
import pandas as pd

df1 = pd.read_excel('people_languge.xlsx')
Oi = df1['Oi'].values[0:5]
Ei = df1['Ei'].values[0:5]
#print(Oi, Ei)
print(chisquare(Oi, Ei))
```

統計假設：

H0：P 國語、P 台語、P 客語、P 原住民族語、P 其他比例是
66.37：37.73：1.5：0.2：0.2

H1：P 國語、P 台語、P 客語、P 原住民族語、P 其他比例不是
66.37：37.73：1.5：0.2：0.2

5. 結論：

```
Power_divergenceResult(statistic=166886753.81923237, pvalue=0.0)
```

P = 0.0 < α = 0.05

X^2 = 166886753.81923237 > $X^2_{0.05}(4)$ = 9.48773

拒絕 H0 =>

在α = 0.05，高雄市各語言分布比例與台灣各語言分布比例無差異的說法是不可以支持的。

第六題

1. 某文化部長想知道 109 年台灣各語言使用的比例是否有差異？

2. Data Source：https://www.thenewslens.com/article/157030

	人數
國語	15,637,592
台語	7,475,980
客語	353,419
原住民族語	47,122
其他	47,122

資料描述：分為每個語言使用的人數及其使用的語言
語言種類有：國語，台語，客語，原住民語，其他

3. 分析方法：卡方適合度檢定

為何採用：X 為語言（類別尺度），Y 為各種語言使用人數（間斷資料）

4. 分析結果：
程式碼（Language_chi.py）：

```
from scipy.stats import chisquare
import pandas as pd

df1 = pd.read_excel('Language.xlsx')
anscount = df1['人數']
print(chisquare(anscount))
```

統計假設：
H0：P 國語 = P 台語 = P 客語 = P 原住民族語 = P 其他
H1：P 國語 != P 台語 != P 客語 != P 原住民族語 != P 其他

5. 結論：

```
Power_divergenceResult(statistic=40220205.5576764, pvalue=0.0)
```

P = 0.0 < α = 0.05

拒絕 H0 =>

在α = 0.05，109 年台灣各語言使用的比例無差異的說法是不可以支持的。

第七題

1. 高雄市政府想知道各月份的運載量是否有差異？

2. Data Source：https://reurl.cc/LppA04

年	月	總運量	日均運量	假日均運量	月台上刷卡日均筆數	車上刷卡日均筆數	售票機日均筆數	補票日均筆數	團體票日均筆數
107	1	275360	8883	15132	1734.2	5495.9	1516.8	9.4	126.3
107	2	413815	14779	20738	3854.9	9039.8	1754.9	11.9	117.7
107	3	209783	6767	13198	1413.2	4487.5	660.7	6.5	199.2
107	4	283651	9455	16656	2183.7	6013.8	944.5	6.3	306.8
107	5	206325	6656	11573	1385.8	4331.5	558.6	7.1	372.7
107	6	213372	7112	11818	1601.2	4673	644.2	7	186.9
107	7	323385	10432	17413	2669.9	6592.6	1011.1	7.5	150.7
107	8	296078	9551	16335	2372.6	6129.6	897.9	5.2	145.6
107	9	218813	7294	12596	1762.7	4751.6	613.3	6.5	159.8
107	10	229454	7402	14376	1712.2	4730.2	617.5	6.5	335.3
107	11	260255	8675	15874	2036.2	5498.7	742	15.2	383.1
107	12	436179	14070	25672	3542.5	9025	1201.3	18.4	283.1
108	1	310193	10006	15375	2713	6238.8	783.2	11.5	259.7
108	2	431570	15413	24853	5475.3	8246.8	1515.5	10.9	164.7
108	3	275955	8902	15843	2661.5	5187.8	721.4	9.2	321.9

資料描述： 輕軌民國 107 年至今每個月份的運載量

3. 分析方法：卡方適合度檢定

為何採用：X 為運載量（類別尺度），Y 為月份（間斷資料）

4. 分析結果：

程式碼（train_chi.py）：

```
from scipy.stats import chisquare
import pandas as pd

df1 = pd.read_excel('train.xlsx')
s = df1['總運量']
print(chisquare(s))
```

統計假設：

H0：各月份運載量分布無差異

H1：各月份運載量分布有差異

5. 結論：

```
Power_divergenceResult(statistic=1976147.1228125456, pvalue=0.0)
```

P = 0.0 < α = 0.05

拒絕 H0 =>

在α = 0.05，各月份運載量分布無差異的說法是不可以支持的。

第八題

1. 某海產業者想知道鮑魚在不同年紀時重量是否有差異？

2. Data Source：https://archive.ics.uci.edu/ml/datasets/Abalone

Gender	Length	Diameter	Height	Whole wei	Shucked w	Viscera w	Shell weig	Rings
M	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15
M	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7
F	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9
M	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10
I	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7
I	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8
F	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20
F	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16
M	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9
F	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19
F	0.525	0.38	0.14	0.6065	0.194	0.1475	0.21	14
M	0.43	0.35	0.11	0.406	0.1675	0.081	0.135	10
M	0.49	0.38	0.135	0.5415	0.2175	0.095	0.19	11
F	0.535	0.405	0.145	0.6845	0.2725	0.171	0.205	10
F	0.47	0.355	0.1	0.4755	0.1675	0.0805	0.185	10
M	0.5	0.4	0.13	0.6645	0.258	0.133	0.24	12
I	0.355	0.28	0.085	0.2905	0.095	0.0395	0.115	7
F	0.44	0.34	0.1	0.451	0.188	0.087	0.13	10
M	0.365	0.295	0.08	0.2555	0.097	0.043	0.1	7
M	0.45	0.32	0.1	0.381	0.1705	0.075	0.115	9

資料描述：年齡斷分為 H (> 20) ，M (19 - 10) ，L (< 10)
重量指的是一顆鮑魚帶殼的重量單位為公斤

3. 分析方法：單因子 ANOVA 檢定

為何採用：X 為三種不同年齡段 (類別尺度)，Y 為鮑魚含殼重量 (連續尺度)

4. 分析結果：
程式碼 (Abalone_anova.py) :

```
import scipy.stats as stats
import pandas as pd

from statsmodels.stats.multicomp import pairwise_tukeyhsd
df1 = pd.read_excel('Abalone.xlsx')

H = []
M = []
L = []

for i in range(len(df1['Rings'])):
```

```
    if df1['Rings'][i] == "H":
        H.append(df1['Whole_weight'][i])
    elif df1['Rings'][i] == "M":
        M.append(df1['Whole_weight'][i])
    else:
        L.append(df1['Whole_weight'][i])

lev = stats.levene(H, M, L)
print('levene test', lev)

if lev[1] > 0.05:
    print('變異數同值')
    ano1 = stats.f_oneway(H, M, L)
    print(ano1)
    if ano1[1] <= 0.05:
        print('事後比較')
        df2 = df1['Rings']
        # print(df2)
        mst = df1['Whole_weight']
        tukey = pairwise_tukeyhsd(endog=mst, groups=df2, alpha=0.05)
        print(tukey.summary())
else:
    krs = stats.kruskal(H, M, L)
    print('變異數異值')
    print(krs)
```

統計假設：

H0 : μL = μM = μH
H1 : μL != μM != μH

5. 結論：

```
levene test LeveneResult(statistic=0.7814680551923499, pvalue=0.4578004982766445)
變異數同值
F_onewayResult(statistic=500.82921872093533, pvalue=1.1135987849371806e-195)
事後比較
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
group1 group2 meandiff p-adj   lower   upper  reject
-----------------------------------------------------
     H      L  -0.6233  0.001 -0.7965   -0.45   True
     H      M  -0.1756 0.0479 -0.3499 -0.0012   True
     L      M   0.4477  0.001  0.4138  0.4816   True
-----------------------------------------------------
```

Lev：p = 0.4578004982766445 > α = 0.05 => 變異數同質
Oneway：p = 1.1135987849371806e-195 < 0.05 => 拒絕 H0
在α = 0.05 下，鮑魚在不同年紀時重量無差異的說法是不可以支持的
事後比較：三種年紀各獨立為一組

第九題

1. 高雄市政府想知道月份對於運載量是否有差異？

2. Data Source：https://reurl.cc/LppA04

年	月	總運量	日均運量	假日均運量	月台上刷卡日均筆數	車上刷卡日均筆數	售票機日均筆數	補票日均筆數	團體票日均筆數
107	1	275360	8883	15132	1734.2	5495.9	1516.8	9.4	126.3
107	2	413815	14779	20738	3854.9	9039.8	1754.9	11.9	117.7
107	3	209783	6767	13198	1413.2	4487.5	660.7	6.5	199.2
107	4	283651	9455	16656	2183.7	6013.8	944.5	6.3	306.8
107	5	206325	6656	11573	1385.8	4331.5	558.6	7.1	372.7
107	6	213372	7112	11818	1601.2	4673	644.2	7	186.9
107	7	323385	10432	17413	2669.9	6592.6	1011.1	7.5	150.7
107	8	296078	9551	16335	2372.6	6129.6	897.9	5.2	145.6
107	9	218813	7294	12596	1762.7	4751.6	613.3	6.5	159.8
107	10	229454	7402	14376	1712.2	4730.2	617.5	6.5	335.3
107	11	260255	8675	15874	2036.2	5498.7	742	15.2	383.1
107	12	436179	14070	25672	3542.5	9025	1201.3	18.4	283.1
108	1	310193	10006	15375	2713	6238.8	783.2	11.5	259.7
108	2	431570	15413	24853	5475.3	8246.8	1515.5	10.9	164.7
108	3	275955	8902	15843	2661.5	5187.8	721.4	9.2	321.9

資料描述：輕軌民國 107 年至今為止每個月份的運載量

3. 分析方法：單因子 ANOVA 檢定

為何採用：X 為月份（類別尺度），Y 為每月總運量（連續尺度）

4. 分析結果：
程式碼（train_onewayanova.py）：

```
import pandas as pd
import scipy.stats as stats
from statsmodels.stats.multicomp import pairwise_tukeyhsd

df1 = pd.read_excel('train.xlsx')
month = df1['月']
s = df1['總運量']

Jan = []
Feb = []
Mar = []
Apr = []
May = []
Jun = []
```

```
Jul = []
Aug = []
Sep = []
Oct = []
Nov = []
Dec = []

for i in range(len(df1['月'])):
    if df1['月'][i] == 1:
        Jan.append(df1['總運量'][i])
    elif df1['月'][i] == 2:
        Feb.append(df1['總運量'][i])
    elif df1['月'][i] == 3:
        Mar.append(df1['總運量'][i])
    elif df1['月'][i] == 4:
        Apr.append(df1['總運量'][i])
    elif df1['月'][i] == 5:
        May.append(df1['總運量'][i])
    elif df1['月'][i] == 6:
        Jun.append(df1['總運量'][i])
    elif df1['月'][i] == 7:
        Jul.append(df1['總運量'][i])
    elif df1['月'][i] == 8:
        Aug.append(df1['總運量'][i])
    elif df1['月'][i] == 9:
        Sep.append(df1['總運量'][i])
    elif df1['月'][i] == 10:
        Oct.append(df1['總運量'][i])
    elif df1['月'][i] == 11:
        Nov.append(df1['總運量'][i])
    else:
        Dec.append(df1['總運量'][i])

#print(Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec)
lev = stats.levene(Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov,
Dec)
print('levene test', lev)
```

```
if lev[1] > 0.05:
    print('變異數同值')
    ano1 = stats.f_oneway(Jan, Feb, Mar, Apr, May, Jun,
                          Jul, Aug, Sep, Oct, Nov, Dec)
    print(ano1)
    if ano1[1] <= 0.05:
        print('事後比較')
        df2 = df1['月']
        # print(df2)
        mst = df1['總運量']
        tukey = pairwise_tukeyhsd(endog=mst, groups=df2, alpha=0.05)
        print(tukey.summary())
else:
    krs = stats.kruskal(Jan, Feb, Mar, Apr, May, Jun,
                        Jul, Aug, Sep, Oct, Nov, Dec)
    print('變異數異值')
    print(krs)
```

統計假設：

H0：月份和運載量無差異

H1：月份和運載量有差異

5. 結論：

2	7	-165246.75	0.3151	-387055.9801	56562.4801	False
2	8	-156086.25	0.3983	-377895.4801	65722.9801	False
2	9	-200907.25	0.1065	-422716.4801	20901.9801	False
2	10	-134568.25	0.5982	-356377.4801	87240.9801	False
2	11	-150052.5	0.4566	-371861.7301	71756.7301	False
2	12	-64409.0833	0.9	-303990.4342	175172.2675	False
3	4	15541.75	0.9	-206267.4801	237350.9801	False
3	5	-41912.75	0.9	-263721.9801	179896.4801	False
3	6	-44714.5	0.9	-266523.7301	177094.7301	False
3	7	24685.0	0.9	-197124.2301	246494.2301	False
3	8	33845.5	0.9	-187963.7301	255654.7301	False
3	9	-10975.5	0.9	-232784.7301	210833.7301	False
3	10	55363.5	0.9	-166445.7301	277172.7301	False
3	11	39879.25	0.9	-181929.9801	261688.4801	False
3	12	125522.6667	0.7644	-114058.6842	365104.0175	False
4	5	-57454.5	0.9	-279263.7301	164354.7301	False
4	6	-60256.25	0.9	-282065.4801	161552.9801	False
4	7	9143.25	0.9	-212665.9801	230952.4801	False
4	8	18303.75	0.9	-203505.4801	240112.9801	False
4	9	-26517.25	0.9	-248326.4801	195291.9801	False
4	10	39821.75	0.9	-181987.4801	261630.9801	False
4	11	24337.5	0.9	-197471.7301	246146.7301	False
4	12	109980.9167	0.8947	-129600.4342	349562.2675	False

5	6	-2801.75	0.9	-224610.9801	219007.4801	False
5	7	66597.75	0.9	-155211.4801	288406.9801	False
5	8	75758.25	0.9	-146050.9801	297567.4801	False
5	9	30937.25	0.9	-190871.9801	252746.4801	False
5	10	97276.25	0.9	-124532.9801	319085.4801	False
5	11	81792.0	0.9	-140017.2301	303601.2301	False
5	12	167435.4167	0.4087	-72145.9342	407016.7675	False
6	7	69399.5	0.9	-152409.7301	291208.7301	False
6	8	78560.0	0.9	-143249.2301	300369.2301	False
6	9	33739.0	0.9	-188070.2301	255548.2301	False
6	10	100078.0	0.9	-121731.2301	321887.2301	False
6	11	84593.75	0.9	-137215.4801	306402.9801	False
6	12	170237.1667	0.3837	-69344.1842	409818.5175	False
7	8	9160.5	0.9	-212648.7301	230969.7301	False
7	9	-35660.5	0.9	-257469.7301	186148.7301	False
7	10	30678.5	0.9	-191130.7301	252487.7301	False
7	11	15194.25	0.9	-206614.9801	237003.4801	False
7	12	100837.6667	0.9	-138743.6842	340419.0175	False

8	9	-44821.0	0.9	-266630.2301	176988.2301	False
8	10	21518.0	0.9	-200291.2301	243327.2301	False
8	11	6033.75	0.9	-215775.4801	227842.9801	False
8	12	91677.1667	0.9	-147904.1842	331258.5175	False
9	10	66339.0	0.9	-155470.2301	288148.2301	False
9	11	50854.75	0.9	-170954.4801	272663.9801	False
9	12	136498.1667	0.6724	-103083.1842	376079.5175	False
10	11	-15484.25	0.9	-237293.4801	206324.9801	False
10	12	70159.1667	0.9	-169422.1842	309740.5175	False
11	12	85643.4167	0.9	-153937.9342	325224.7675	False

事後比較

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
1	2	72976.25	0.9	-148832.9801	294785.4801	False
1	3	-116955.5	0.7577	-338764.7301	104853.7301	False
1	4	-101413.75	0.8984	-323222.9801	120395.4801	False
1	5	-158868.25	0.3718	-380677.4801	62940.9801	False
1	6	-161670.0	0.3459	-383479.2301	60139.2301	False
1	7	-92270.5	0.9	-314079.7301	129538.7301	False
1	8	-83110.0	0.9	-304919.2301	138699.2301	False
1	9	-127931.0	0.6583	-349740.2301	93878.2301	False
1	10	-61592.0	0.9	-283401.2301	160217.2301	False
1	11	-77076.25	0.9	-298885.4801	144732.9801	False
1	12	8567.1667	0.9	-231014.1842	248148.5175	False
2	3	-189931.75	0.1531	-411740.9801	31877.4801	False
2	4	-174390.0	0.2452	-396199.2301	47419.2301	False
2	5	-231844.5	0.034	-453653.7301	-10035.2699	True
2	6	-234646.25	0.0304	-456455.4801	-12837.0199	True

Lev：p = 0.7909309084131676 > α = 0.05 => 變異數同質

Oneway：p = 0.02435079091858752 < 0.05 => 拒絕 H0

在 α = 0.05 下，輕軌在不同月份時總運量無差異的說法是不可以支持的。

事後比較：2 月與 5 月、2 月與 6 月，非為一組；其餘的組合則反之

第十題

1. 我們欲檢視飲料杯數是否隨著氣溫增加而遞增，即在顯著水準 0.05 下，檢定斜率 β1 是否等於零？

2. Data Source：https://reurl.cc/bkkGRr

date	quantity	temperature
9/18	277	27
9/19	243	28
9/20	231	27
9/21	212	28
9/22	215	29
9/23	231	31
9/24	208	30
9/25	217	32
9/26	233	31
11/22	98	20
11/23	103	19
11/24	116	21
11/25	121	22
11/26	127	21
11/27	137	24
11/28	148	23
11/29	134	23
11/30	96	17

資料描述：冬季、夏季的氣溫&飲料銷售杯數

3. 分析方法：回歸分析

為何採用：X 為氣溫（連續尺度），Y 為飲料銷售杯數（連續尺度）

4. 分析結果：
程式碼（Drink_R.py）：

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
import pandas as pd

df = pd.read_excel('drink.xlsx')
data_x = df['temperature']
data_y = df['quantity']

# 分割 trian/test
```

```
train_x, test_x, train_y, test_y = train_test_split(data_x,
                                                    data_y,
                                                    test_size=0.25,
                                                    random_state=1)

# 氣溫對杯數的散佈圖

plt.scatter(train_x, train_y, facecolor='None', edgecolor='k', alpha=0.3)
plt.show()

train_x.values.reshape(-1, 1)

train_x_reshape = train_x.values.reshape(-1, 1)

model = LinearRegression()
model.fit(train_x_reshape, train_y)

print(f"Coefficient:{model.coef_}")
print(f"Beta0:{model.intercept_}")
print(f"R2:{model.score(train_x_reshape, train_y)}")

data_x_reshape = data_x.values.reshape(-1, 1)

xp = np.linspace(data_x_reshape.min(), data_x_reshape.max(), 70)
xp = xp.reshape(-1, 1)
pred_plot = model.predict(xp)

plt.title('LinearRegression')
plt.xlabel('Temperature')
plt.ylabel('Quantity')

plt.scatter(data_x_reshape, data_y, facecolor='None', edgecolor='k',
alpha=0.8)
plt.plot(xp, pred_plot)
plt.show()

test_x_reshape = test_x.values.reshape(-1, 1)
```

```
pred = model.predict(test_x_reshape)
print("predict")
for i in range(5):
    print("true:", test_y.values[i])
    print("prediction:", pred[i])
```

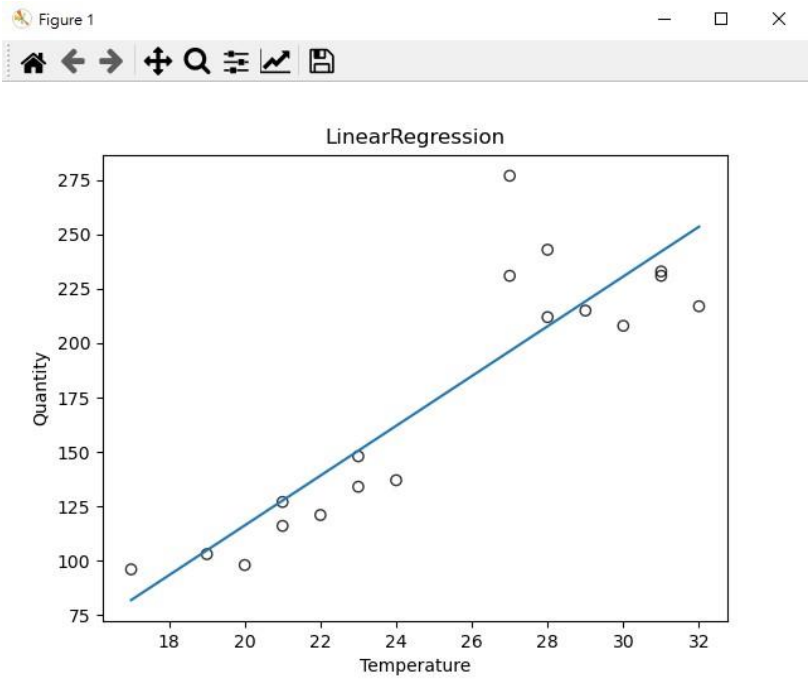
統計假設：

H0：氣溫與飲料杯數不呈現性關係

H1：氣溫與飲料杯數呈現性關係

5. 結論：

```
Coefficient:[11.44775]
Beta0:-112.74025000000006
R2:0.7920402517980176
```

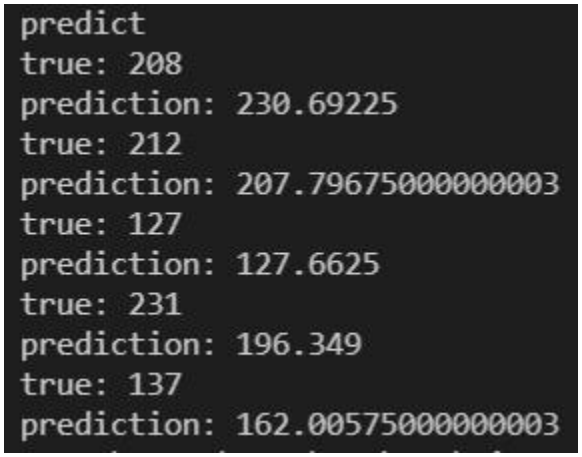


$\hat{Y}$=-112.74025+11.44775X

R-squard：0.7920402517980176，表示此模型在 X->Y 的解釋能力為 79.2%

Coefficient：11.44775 > 0 => 正相關

預測：



氣溫對飲料店的銷售飲料杯數呈正線性關係