整理候選人得票表格

透過網路抓取資訊並把表格整理成Tidy Form

In [1]:
```
from urllib.parse import quote_plus
from string import ascii_uppercase
import pandas as pd
```

In [2]:
```
file_name = '總統-A05-4-候選人得票數一覽表-各投開票所(臺北市).xls'
file_url = quote_plus(file_name)
spreadsheet_url = "https://taiwan-election-data.s3-ap-northeast-1.amazonaws.com/presidential_2020/{}".format(file_url)
```

In [3]:
```
presidential = pd.read_excel(spreadsheet_url, skiprows=[0, 1, 3, 4], thousands=',')
presidential
```

Out[3]:

	Unnamed: 0	Unnamed: 1	Unnamed: 2	(1)\n宋楚瑜\n余湘	(2)\n韓國瑜\n張善政	(3)\n蔡英文\n賴清德	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11	Unnamed: 12	Unnamed: 13
0	總　計	NaN	NaN	70769	685830	875854	1632453	21381	1653834	143	1653977	513287	2167264	76.3098
1	松山區	NaN	NaN	5436	55918	64207	125561	1762	127323	2	127325	37329	164654	77.3276
2	NaN	莊敬里	573.0	36	391	429	856	14	870	0	870	230	1100	79.0909
3	NaN	莊敬里	574.0	46	382	438	866	12	878	0	878	259	1137	77.2208
4	NaN	莊敬里	575.0	48	393	389	830	22	852	0	852	262	1114	76.4811
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1736	NaN	泉源里	156.0	48	372	721	1141	8	1149	0	1149	448	1597	71.9474
1737	NaN	湖山里	157.0	25	219	344	588	5	593	0	593	212	805	73.6646
1738	NaN	湖山里	158.0	23	191	282	496	7	503	0	503	188	691	72.7931
1739	NaN	大屯里	159.0	34	195	542	771	10	781	0	781	300	1081	72.2479
1740	NaN	湖田里	160.0	27	225	350	602	4	606	0	606	229	835	72.5749

1741 rows × 14 columns

In [4]:
```
n_cols = presidential.columns.size
n_candidate = n_cols-11
idvars=['town', 'village', 'office']
candidate = list(presidential.columns[3:(3+n_candidate)])
office = list(ascii_uppercase[:8])
columns = idvars+candidate+office
presidential.columns = columns
presidential
```

Out[4]:

	town	village	office	(1)\n宋楚瑜\n余湘	(2)\n韓國瑜\n張善政	(3)\n蔡英文\n賴清德	A	B	C	D	E	F	G	H
0	總　計	NaN	NaN	70769	685830	875854	1632453	21381	1653834	143	1653977	513287	2167264	76.3098
1	松山區	NaN	NaN	5436	55918	64207	125561	1762	127323	2	127325	37329	164654	77.3276
2	NaN	莊敬里	573.0	36	391	429	856	14	870	0	870	230	1100	79.0909
3	NaN	莊敬里	574.0	46	382	438	866	12	878	0	878	259	1137	77.2208
4	NaN	莊敬里	575.0	48	393	389	830	22	852	0	852	262	1114	76.4811
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1736	NaN	泉源里	156.0	48	372	721	1141	8	1149	0	1149	448	1597	71.9474
1737	NaN	湖山里	157.0	25	219	344	588	5	593	0	593	212	805	73.6646
1738	NaN	湖山里	158.0	23	191	282	496	7	503	0	503	188	691	72.7931
1739	NaN	大屯里	159.0	34	195	542	771	10	781	0	781	300	1081	72.2479
1740	NaN	湖田里	160.0	27	225	350	602	4	606	0	606	229	835	72.5749

1741 rows × 14 columns

In [5]:
```
presidential_filna = presidential.fillna(method='ffill')
```

In [6]:
```
presidential_dropna = presidential_filna.dropna()
presidential_dropna
```

Out[6]:

	town	village	office	(1)\n宋楚瑜\n余湘	(2)\n韓國瑜\n張善政	(3)\n蔡英文\n賴清德	A	B	C	D	E	F	G	H
2	松山區	莊敬里	573.0	36	391	429	856	14	870	0	870	230	1100	79.0909
3	松山區	莊敬里	574.0	46	382	438	866	12	878	0	878	259	1137	77.2208
4	松山區	莊敬里	575.0	48	393	389	830	22	852	0	852	262	1114	76.4811
5	松山區	莊敬里	576.0	43	389	462	894	14	908	0	908	271	1179	77.0144
6	松山區	東榮里	577.0	38	431	545	1014	18	1032	0	1032	272	1304	79.1411
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1736	北投區	泉源里	156.0	48	372	721	1141	8	1149	0	1149	448	1597	71.9474
1737	北投區	湖山里	157.0	25	219	344	588	5	593	0	593	212	805	73.6646
1738	北投區	湖山里	158.0	23	191	282	496	7	503	0	503	188	691	72.7931
1739	北投區	大屯里	159.0	34	195	542	771	10	781	0	781	300	1081	72.2479
1740	北投區	湖田里	160.0	27	225	350	602	4	606	0	606	229	835	72.5749

1739 rows × 14 columns

In [7]:
```
df_presidential = pd.melt(presidential_dropna, id_vars=idvars, var_name='candidate', value_name='vote')
df_presidential
```

Out[7]:

	town	village	office	candidate	vote
0	松山區	莊敬里	573.0	(1)\n宋楚瑜\n余湘	36.0000
1	松山區	莊敬里	574.0	(1)\n宋楚瑜\n余湘	46.0000
2	松山區	莊敬里	575.0	(1)\n宋楚瑜\n余湘	48.0000
3	松山區	莊敬里	576.0	(1)\n宋楚瑜\n余湘	43.0000
4	松山區	東榮里	577.0	(1)\n宋楚瑜\n余湘	38.0000
...	...	...	...	...	...
19124	北投區	泉源里	156.0	H	71.9474
19125	北投區	湖山里	157.0	H	73.6646
19126	北投區	湖山里	158.0	H	72.7931
19127	北投區	大屯里	159.0	H	72.2479
19128	北投區	湖田里	160.0	H	72.5749

19129 rows × 5 columns