In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

In [2]:
```
csv_in = 'cs3-mid-2-1.csv'
df1 = pd.read_csv(csv_in, skiprows=0, sep=',', header=0)
```

In [3]:
```
csv_in2 = 'cs3-mid-2-2.csv'
df2 = pd.read_csv(csv_in2, skiprows=0, sep=',', header=0)
```

In [4]:
```
print(df1.shape)
print(df1.info())
display(df1.head())
```

```
(81, 7)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 81 entries, 0 to 80
Data columns (total 7 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   ID      81 non-null     object
 1   b1      81 non-null     float64
 2   b2      80 non-null     float64
 3   b3      81 non-null     float64
 4   b4      80 non-null     float64
 5   b5      81 non-null     float64
 6   c1      81 non-null     object
dtypes: float64(5), object(2)
memory usage: 4.6+ KB
None
```

	ID	b1	b2	b3	b4	b5	c1
0	ID00	-1.04	0.22	-0.50	-0.60	-0.92	C
1	ID01	-1.27	NaN	-0.42	-0.79	-1.42	C
2	ID02	-0.58	1.00	-0.32	-0.58	-1.54	C
3	ID03	-0.70	-1.85	0.17	0.15	1.61	A
4	ID04	2.34	0.68	1.54	1.05	0.57	B

(1)

In [5]:
```
display(df1[df1.duplicated(keep=False)])
```

	ID	b1	b2	b3	b4	b5	c1
3	ID03	-0.7	-1.85	0.17	0.15	1.61	A
80	ID03	-0.7	-1.85	0.17	0.15	1.61	A

(2)

In [6]:
```
df1m = df1.drop_duplicates().reset_index(drop=True)
```

In [7]:
```
print(df1m.shape)
```

```
(80, 7)
```

(3)

80

(4)

In [8]:
```
print( df1m.isna().sum(axis=0) )
```

```
ID    0
b1    0
b2    1
b3    0
b4    1
b5    0
c1    0
dtype: int64
```

(5)

b2, b4

(6)

In [9]:
```
display( df1m[df1m.isnull().any(axis=1)] )
```

	ID	b1	b2	b3	b4	b5	c1
1	ID01	-1.27	NaN	-0.42	-0.79	-1.42	C
20	ID20	-1.95	-2.6	-1.12	NaN	1.88	A

(7)

In [10]:
```
df1m2 = df1m.dropna().reset_index(drop=True)
```

(8)

In [11]:
```
display( df2['c2'].value_counts() )
```

```
c2
X    23
Y    18
Z    14
W    14
y     1
Name: count, dtype: int64
```

(9)

23

(10)

In [12]:
```
df2['c2']=df2['c2'].replace('y', 'Y')
```

(11)

In [13]:
```
df3=pd.merge(df1m2, df2, left_on='ID', right_on='idx', how='inner')
```

(12)

In [14]:
```
df3.to_csv('mid-p2-out.csv', index=False)
```

In [15]:
```
print(df3.shape)
```

```
(68, 9)
```

(13)(14)

68, 9

(15)

In [16]:
```
display( pd.crosstab(df3['c1'], df3['c2']) )
```

c2	W	X	Y	Z
c1				
A	4	9	8	1
B	3	8	7	7
C	5	6	4	6

(16)

8