CS3-mid-p2

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

In [2]:
```
csv_in1 = 'mid-p2-1.csv'
csv_in2 = 'mid-p2-2.csv'
```

In [3]:
```
s1 = pd.read_csv(csv_in1, sep=',', skiprows=0, header=0)
print(s1.shape)
print(s1.info())
display(s1.head())
```

```
(58, 5)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 58 entries, 0 to 57
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   h1      58 non-null     int64  
 1   h2      57 non-null     float64
 2   h3      55 non-null     float64
 3   h4      58 non-null     object 
 4   h5      58 non-null     object 
dtypes: float64(2), int64(1), object(2)
memory usage: 2.4+ KB
None
```

	h1	h2	h3	h4	h5
0	2	7.0	5.0	n	E
1	18	7.0	6.0	k	C
2	6	2.0	13.0	f	E
3	11	1.0	9.0	b	E
4	10	7.0	20.0	k	E

In [4]:
```
s2 = pd.read_csv(csv_in2, sep=',', skiprows=0, header=0)
print(s2.shape)
print(s2.info())
display(s2.head())
```

```
(14, 2)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14 entries, 0 to 13
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   alpha   14 non-null     object
 1   spell   14 non-null     object
dtypes: object(2)
memory usage: 352.0+ bytes
None
```

	alpha	spell
0	A	Able
1	B	Baker
2	C	Charlie
3	D	Delta
4	E	Echo

In [5]:
```
display( s1[ s1.duplicated(keep=False) ] )  # (1)
s1d = s1.drop_duplicates().reset_index(drop=True)  # (2)
print(s1d.shape)
```

	h1	h2	h3	h4	h5
1	18	7.0	6.0	k	C
4	10	7.0	20.0	k	E
10	18	7.0	6.0	k	C
15	18	7.0	6.0	k	C
22	18	7.0	6.0	k	C
31	10	7.0	20.0	k	E

```
(54, 5)
```

(3) 54

In [6]:
```
print(s1d.isna().sum())  # (4)
display(s1d[ s1d.isna().any(axis=1) ])  # (6)
s1d2 = s1d.dropna().reset_index(drop=True)  # (7)
s1d2['h4'] = s1d2['h4'].replace('f', 'fff')  # (8)
```

```
h1    0
h2    1
h3    3
h4    0
h5    0
dtype: int64
```

	h1	h2	h3	h4	h5
5	18	NaN	2.0	b	M
22	1	10.0	NaN	b	M
31	8	12.0	NaN	k	E
32	5	10.0	NaN	n	C

(5) 0

In [7]:
```
s1d2['h2'] = s1d2['h2'].astype('int')  # (9)
s1d2['h3'] = s1d2['h3'].astype('int')
```

In [8]:
```
s3 = pd.merge(s1d2, s2, how='left', left_on='h5', right_on='alpha')  # (10)
```

In [9]:

```
s3.to_csv('mid-p2-out.csv', index=False)  # (11)
print(s3.shape)
```

```
(50, 7)
```

(12) 50

(13) 7