Load the libraries

In [77]:
```
import numpy as np
import pandas as pd
```

Load the data

In [78]:
```
df = pd.read_excel (r'/Users/ghosebishwajit/pCloud Drive/dhsdf.xlsx')
print (df)
```

```
       age   cage division  place           edu electricity toilet   wealth  \
0       26  25-29  barisal  rural       primary          no     no  poorest
1       41  40-44  barisal  rural       primary          no     no   middle
2       38  35-39  barisal  rural       primary          no     no  poorest
3       18  15-19  barisal  rural     secondary         NaN    NaN  poorest
4       23  20-24  barisal  rural  no education          no    yes  poorest
...    ...    ...      ...    ...           ...         ...    ...      ...
17858   20  20-24   sylhet  urban       primary         NaN    NaN   middle
17859   22  20-24   sylhet  urban     secondary         yes     no  richest
17860   26  25-29   sylhet  urban        higher         yes     no  richest
17861   49  45-49   sylhet  urban  no education         yes     no  richest
17862   30  30-34   sylhet  urban     secondary         yes     no  richest

       parity  anc     bmi  stunt  wasting  underwight  intercourse      bm  \
0           2    2  2096.0 -104.0   -252.0      -234.0          2.0   Lower
1           4  NaN  1656.0    NaN      NaN         NaN          6.0   Lower
2           2  NaN  1857.0    NaN      NaN         NaN          3.0   Lower
3           0  NaN  1738.0    NaN      NaN         NaN          3.0   Lower
4           2    1     NaN    NaN      NaN         NaN          0.0     NaN
...       ...  ...     ...    ...      ...         ...          ...     ...
17858       1    8  2122.0 -179.0   -226.0      -151.0          7.0   Lower
17859       1  NaN  2351.0    NaN      NaN         NaN          0.0  Higher
17860       2    7  2360.0    NaN      NaN         NaN          6.0  Higher
17861       4  NaN  1856.0    NaN      NaN         NaN          2.0   Lower
17862       2  NaN  2410.0    NaN      NaN         NaN          0.0  Higher

           st    wst    und
0      Higher  Lower  Lower
1         NaN    NaN    NaN
2         NaN    NaN    NaN
3         NaN    NaN    NaN
4         NaN    NaN    NaN
...       ...    ...    ...
17858   Lower  Lower  Lower
17859     NaN    NaN    NaN
17860     NaN    NaN    NaN
17861     NaN    NaN    NaN
17862     NaN    NaN    NaN

[17863 rows x 19 columns]
```

Print the names of the columns

In [79]:
```
df.columns
```

Out[79]:
```
Index(['age', 'cage', 'division', 'place', 'edu', 'electricity', 'toilet',
       'wealth', 'parity', 'anc', 'bmi', 'stunt', 'wasting', 'underwight',
```

```
       'intercourse', 'bm', 'st', 'wst', 'und'],
      dtype='object')
```

In [80]:
```
list(df)
```

Out[80]:
```
['age',
 'cage',
 'division',
 'place',
 'edu',
 'electricity',
 'toilet',
 'wealth',
 'parity',
 'anc',
 'bmi',
 'stunt',
 'wasting',
 'underwight',
 'intercourse',
 'bm',
 'st',
 'wst',
 'und']
```

describe the df

In [81]:
```
df.describe()
```

Out[81]:

	age	parity	bmi	stunt	wasting	underwight	intercourse
count	17863.000000	17863.000000	17683.000000	6171.000000	6171.000000	6171.000000	17859.000000
mean	31.015955	2.450428	2230.798846	-132.551774	-158.040188	-97.134014	4.663027
std	9.220803	1.750567	413.503903	127.162210	112.334859	103.083652	4.915970
min	15.000000	0.000000	1206.000000	-592.000000	-551.000000	-398.000000	0.000000
25%	23.000000	1.000000	1920.000000	-212.000000	-235.500000	-165.000000	1.000000
50%	30.000000	2.000000	2184.000000	-136.000000	-170.000000	-107.000000	4.000000
75%	38.000000	3.000000	2486.000000	-55.000000	-93.000000	-37.000000	6.000000
max	49.000000	15.000000	5088.000000	553.000000	522.000000	497.000000	74.000000

Print data types

In [82]:
```
df.dtypes
```

Out[82]:
```
age              int64
cage            object
division        object
place           object
edu             object
electricity     object
toilet          object
wealth          object
parity           int64
anc             object
bmi            float64
stunt          float64
```

```
wasting         float64
underwight      float64
intercourse     float64
bm               object
st               object
wst              object
und              object
dtype: object
```

Keep only numeric columns

In [83]:
```
df1 = df.select_dtypes(include='number')
```

In [84]:
```
df1
```

Out[84]:

	age	parity	bmi	stunt	wasting	underwight	intercourse
0	26	2	2096.0	-104.0	-252.0	-234.0	2.0
1	41	4	1656.0	NaN	NaN	NaN	6.0
2	38	2	1857.0	NaN	NaN	NaN	3.0
3	18	0	1738.0	NaN	NaN	NaN	3.0
4	23	2	NaN	NaN	NaN	NaN	0.0
...	...	...	...	...	...	...	...
17858	20	1	2122.0	-179.0	-226.0	-151.0	7.0
17859	22	1	2351.0	NaN	NaN	NaN	0.0
17860	26	2	2360.0	NaN	NaN	NaN	6.0
17861	49	4	1856.0	NaN	NaN	NaN	2.0
17862	30	2	2410.0	NaN	NaN	NaN	0.0

17863 rows × 7 columns

Keep only string columns

In [85]:
```
df2 = df.select_dtypes(include='object')
```

In [86]:
```
df2
```

Out[86]:

	cage	division	place	edu	electricity	toilet	wealth	anc	bm	st	wst	und
0	25-29	barisal	rural	primary	no	no	poorest	2	Lower	Higher	Lower	Lower
1	40-44	barisal	rural	primary	no	no	middle	NaN	Lower	NaN	NaN	NaN
2	35-39	barisal	rural	primary	no	no	poorest	NaN	Lower	NaN	NaN	NaN
3	15-19	barisal	rural	secondary	NaN	NaN	poorest	NaN	Lower	NaN	NaN	NaN
4	20-24	barisal	rural	no education	no	yes	poorest	1	NaN	NaN	NaN	NaN

17858	20-24	sylhet	urban	primary	NaN	NaN	middle	8	Lower	Lower	Lower	Lower
17859	20-24	sylhet	urban	secondary	yes	no	richest	NaN	Higher	NaN	NaN	NaN
17860	25-29	sylhet	urban	higher	yes	no	richest	7	Higher	NaN	NaN	NaN
17861	45-49	sylhet	urban	no education	yes	no	richest	NaN	Lower	NaN	NaN	NaN
17862	30-34	sylhet	urban	secondary	yes	no	richest	NaN	Higher	NaN	NaN	NaN

17863 rows × 12 columns

Keeping specific columns by names

In [87]:
```
df3 = df[["age","place","bmi", "edu"]]
```

In [88]:
```
df3
```

Out[88]:

	age	place	bmi	edu
0	26	rural	2096.0	primary
1	41	rural	1656.0	primary
2	38	rural	1857.0	primary
3	18	rural	1738.0	secondary
4	23	rural	NaN	no education
...	...	...	...	...
17858	20	urban	2122.0	primary
17859	22	urban	2351.0	secondary
17860	26	urban	2360.0	higher
17861	49	urban	1856.0	no education
17862	30	urban	2410.0	secondary

17863 rows × 4 columns

Or by column position

In [89]:
```
df3 = df.iloc[:,[1,3,5,6]]
```

In [90]:
```
df3
```

Out[90]:

	cage	place	electricity	toilet
0	25-29	rural	no	no
1	40-44	rural	no	no

2	35-39	rural	no	no
3	15-19	rural	NaN	NaN
4	20-24	rural	no	yes
...	...	...	...	...
17858	20-24	urban	NaN	NaN
17859	20-24	urban	yes	no
17860	25-29	urban	yes	no
17861	45-49	urban	yes	no
17862	30-34	urban	yes	no

17863 rows × 4 columns

Select columns by strings contained in the column names

In [91]:
```
df4 = df.filter(regex='ag|ed|we')
```

In [92]:
```
df4
```

Out[92]:

	age	cage	edu	wealth
0	26	25-29	primary	poorest
1	41	40-44	primary	middle
2	38	35-39	primary	poorest
3	18	15-19	secondary	poorest
4	23	20-24	no education	poorest
...	...	...	...	...
17858	20	20-24	primary	middle
17859	22	20-24	secondary	richest
17860	26	25-29	higher	richest
17861	49	45-49	no education	richest
17862	30	30-34	secondary	richest

17863 rows × 4 columns

or

In [93]:
```
df4 = df[df.columns[df.columns.str.contains('ed|we|pl|st')]]
```

In [94]:
```
df4
```

Out[94]:

	place	edu	wealth	stunt	wasting	st	wst
0	rural	primary	poorest	-104.0	-252.0	Higher	Lower

1	rural	primary	middle	NaN	NaN	NaN	NaN
2	rural	primary	poorest	NaN	NaN	NaN	NaN
3	rural	secondary	poorest	NaN	NaN	NaN	NaN
4	rural	no education	poorest	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...
17858	urban	primary	middle	-179.0	-226.0	Lower	Lower
17859	urban	secondary	richest	NaN	NaN	NaN	NaN
17860	urban	higher	richest	NaN	NaN	NaN	NaN
17861	urban	no education	richest	NaN	NaN	NaN	NaN
17862	urban	secondary	richest	NaN	NaN	NaN	NaN

17863 rows × 7 columns

Drop columns that match a certain string

In [95]:
```
df5 = df.drop(df.filter(like ='ag|st|ws',axis=1).columns,axis=1)
```

In [96]:
```
df5
```

Out[96]:

	age	cage	division	place	edu	electricity	toilet	wealth	parity	anc	bmi	stunt	was
0	26	25-29	barisal	rural	primary	no	no	poorest	2	2	2096.0	-104.0	-2
1	41	40-44	barisal	rural	primary	no	no	middle	4	NaN	1656.0	NaN	
2	38	35-39	barisal	rural	primary	no	no	poorest	2	NaN	1857.0	NaN	
3	18	15-19	barisal	rural	secondary	NaN	NaN	poorest	0	NaN	1738.0	NaN	
4	23	20-24	barisal	rural	no education	no	yes	poorest	2	1	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	
17858	20	20-24	sylhet	urban	primary	NaN	NaN	middle	1	8	2122.0	-179.0	-2
17859	22	20-24	sylhet	urban	secondary	yes	no	richest	1	NaN	2351.0	NaN	
17860	26	25-29	sylhet	urban	higher	yes	no	richest	2	7	2360.0	NaN	
17861	49	45-49	sylhet	urban	no education	yes	no	richest	4	NaN	1856.0	NaN	
17862	30	30-34	sylhet	urban	secondary	yes	no	richest	2	NaN	2410.0	NaN	

17863 rows × 19 columns

Subsetting based on one condition

In [97]:
```
df7 = df.loc[(df['age']>=30)]
```

In [98]:
```
df7
```

Out[98]:

	age	cage	division	place	edu	electricity	toilet	wealth	parity	anc	bmi	stunt	was
1	41	40-44	barisal	rural	primary	no	no	middle	4	NaN	1656.0	NaN	
2	38	35-39	barisal	rural	primary	no	no	poorest	2	NaN	1857.0	NaN	
5	30	30-34	barisal	rural	no education	no	no	poorest	2	NaN	1876.0	NaN	
7	34	30-34	barisal	rural	primary	no	no	poorer	3	6	2066.0	-61.0	-1
8	44	40-44	barisal	rural	primary	no	yes	poorer	4	NaN	2150.0	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	
17854	36	35-39	sylhet	urban	primary	NaN	NaN	middle	2	NaN	2512.0	NaN	
17856	30	30-34	sylhet	urban	primary	yes	yes	richer	3	NaN	2240.0	-481.0	-2
17857	36	35-39	sylhet	urban	no education	yes	yes	middle	4	NaN	1601.0	NaN	
17861	49	45-49	sylhet	urban	no education	yes	no	richest	4	NaN	1856.0	NaN	
17862	30	30-34	sylhet	urban	secondary	yes	no	richest	2	NaN	2410.0	NaN	

9336 rows × 19 columns

Subsetting based on multiple conditions

In [99]:
```
df8 = df.loc[(df['age']>=30) & (df['parity']< 5) & (df['und']=='Lower')]
```

In [100…
```
df8
```

Out[100]:

	age	cage	division	place	edu	electricity	toilet	wealth	parity	anc	bmi	stunt	wa
7	34	30-34	barisal	rural	primary	no	no	poorer	3	6	2066.0	-61.0	-
14	35	35-39	barisal	rural	primary	no	no	middle	4	NaN	1947.0	-196.0	-2
21	34	30-34	barisal	rural	no education	no	no	poorer	3	NaN	1844.0	-368.0	-
23	31	30-34	barisal	rural	primary	yes	no	poorer	3	NaN	2385.0	-225.0	-2
39	31	30-34	barisal	rural	primary	yes	no	poorer	2	2	1850.0	-52.0	-
...	...	...	...	...	...	...	...	...	...	...	...	...	
17757	38	35-39	sylhet	urban	no education	no	no	middle	4	2	1615.0	-80.0	-

17769	30	30-34	sylhet	urban	secondary	yes	no	richest	3	2	3493.0	-84.0	-
17810	39	35-39	sylhet	urban	no education	yes	yes	poorer	2	NaN	1917.0	-304.0	-
17830	35	35-39	sylhet	urban	secondary	yes	no	richest	3	NaN	3125.0	-27.0	-
17850	30	30-34	sylhet	urban	primary	yes	yes	middle	3	NaN	2148.0	-50.0	-

626 rows × 19 columns

In [101…
```
df9 =  df.loc[(df['age']>=30) & (df['parity']< 5) & (df['division']=='barisal') & df['un
```

In [102…
```
df9
```

Out[102]:

	age	cage	division	place	edu	electricity	toilet	wealth	parity	anc	bmi	stunt	was
7	34	30-34	barisal	rural	primary	no	no	poorer	3	6	2066.0	-61.0	-1
14	35	35-39	barisal	rural	primary	no	no	middle	4	NaN	1947.0	-196.0	-2
21	34	30-34	barisal	rural	no education	no	no	poorer	3	NaN	1844.0	-368.0	-3
23	31	30-34	barisal	rural	primary	yes	no	poorer	3	NaN	2385.0	-225.0	-2
39	31	30-34	barisal	rural	primary	yes	no	poorer	2	2	1850.0	-52.0	-2
...	...	...	...	...	...	...	...	...	...	...	...	...	
2010	31	30-34	barisal	urban	primary	yes	no	richer	2	2	2097.0	-209.0	-2
2070	36	35-39	barisal	urban	higher	yes	no	richest	3	NaN	2932.0	68.0	-1
2079	30	30-34	barisal	urban	secondary	yes	no	richest	3	1	2206.0	-166.0	-2
2095	31	30-34	barisal	urban	primary	yes	yes	middle	4	1	2815.0	-10.0	-1
2134	31	30-34	barisal	urban	secondary	yes	no	poorer	3	NaN	1854.0	-201.0	-2

79 rows × 19 columns

In [103…
```
df.isnull()
```

Out[103]:

	age	cage	division	place	edu	electricity	toilet	wealth	parity	anc	bmi	stunt	wasting
0	False	False	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	True	False	True	True
2	False	False	False	False	False	False	False	False	False	True	False	True	True
3	False	False	False	False	False	True	True	False	False	True	False	True	True
4	False	False	False	False	False	False	False	False	False	False	True	True	True
...	...	...	...	...	...	...	...	...	...	...	...	...	...

17858	False	False	False	False	False	True	True	False	False	False	False	False	False
17859	False	False	False	False	False	False	False	False	False	True	False	True	True
17860	False	False	False	False	False	False	False	False	False	False	False	True	True
17861	False	False	False	False	False	False	False	False	False	True	False	True	True
17862	False	False	False	False	False	False	False	False	False	True	False	True	True

17863 rows × 19 columns