Data in Motion Pandas Challenge Week 7

import

In [1]:

```
import pandas as pd
import numpy as np
```

In [2]:

```
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data'

wine = pd.read_csv(url)
```

In [3]:

```
wine.head()
```

Out[3]:

	1	14.23	1.71	2.43	15.6	127	2.8	3.06	.28	2.29	5.64	1.04	3.92	1065
0	1	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.40	1050
1	1	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.81	5.68	1.03	3.17	1185
2	1	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.18	7.80	0.86	3.45	1480
3	1	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	1.82	4.32	1.04	2.93	735
4	1	14.20	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.75	1.05	2.85	1450

In [4]:

```
wine.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 177 entries, 0 to 176
Data columns (total 14 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   1       177 non-null    int64  
 1   14.23   177 non-null    float64
 2   1.71    177 non-null    float64
 3   2.43    177 non-null    float64
 4   15.6    177 non-null    float64
 5   127     177 non-null    int64  
 6   2.8     177 non-null    float64
 7   3.06    177 non-null    float64
 8   .28     177 non-null    float64
 9   2.29    177 non-null    float64
 10  5.64    177 non-null    float64
 11  1.04    177 non-null    float64
 12  3.92    177 non-null    float64
 13  1065    177 non-null    int64  
dtypes: float64(11), int64(3)
memory usage: 19.5 KB
```

In [5]:

```
df=wine.copy()
```

Delete the first, fourth, seventh, nineth, eleventh, thirteenth and fourteenth columns.

In [6]:

```
df.drop(df.columns[[0,3,6,8,10,12,13]],axis=1,inplace=True)
```

In [7]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 177 entries, 0 to 176
Data columns (total 7 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   14.23   177 non-null    float64
 1   1.71    177 non-null    float64
 2   15.6    177 non-null    float64
 3   127     177 non-null    int64  
 4   3.06    177 non-null    float64
 5   2.29    177 non-null    float64
 6   1.04    177 non-null    float64
dtypes: float64(6), int64(1)
memory usage: 9.8 KB
```

In [8]:
```
df.head()
```

Out[8]:

	14.23	1.71	15.6	127	3.06	2.29	1.04
0	13.20	1.78	11.2	100	2.76	1.28	1.05
1	13.16	2.36	18.6	101	3.24	2.81	1.03
2	14.37	1.95	16.8	113	3.49	2.18	0.86
3	13.24	2.59	21.0	118	2.69	1.82	1.04
4	14.20	1.76	15.2	112	3.39	1.97	1.05

Assign the columns as below:

- alcohol
- malic_acid
- alcalinity_of_ash
- magnesium
- flavanoids
- proanthocyanins
- hue

In [9]:
```
df.columns=['alcohol','malic_acid','alcalinity_of_ash','magnesium','flavanoids','proanth
```

In [10]:
```
df.head()
```

Out[10]:

	alcohol	malic_acid	alcalinity_of_ash	magnesium	flavanoids	proanthocyanins	hue
0	13.20	1.78	11.2	100	2.76	1.28	1.05
1	13.16	2.36	18.6	101	3.24	2.81	1.03
2	14.37	1.95	16.8	113	3.49	2.18	0.86
3	13.24	2.59	21.0	118	2.69	1.82	1.04
4	14.20	1.76	15.2	112	3.39	1.97	1.05

Set the values of the first 3 rows in the alcohol column as NaN

In [11]:
```
df.loc[:2,'alcohol']=np.nan
```

In [12]:
```
df.head()
```

Out[12]:

	alcohol	malic_acid	alcalinity_of_ash	magnesium	flavanoids	proanthocyanins	hue

0	NaN	1.78	11.2	100	2.76	1.28	1.05
1	NaN	2.36	18.6	101	3.24	2.81	1.03
2	NaN	1.95	16.8	113	3.49	2.18	0.86
3	13.24	2.59	21.0	118	2.69	1.82	1.04
4	14.20	1.76	15.2	112	3.39	1.97	1.05

Now set the value of the rows 3 and 4 of the magnesium column as NaN

In [13]:

```
df.magnesium.iloc[2:4]=np.nan
df.head()
```

```
/tmp/ipykernel_8274/2826667318.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_
guide/indexing.html#returning-a-view-versus-a-copy
  df.magnesium.iloc[2:4]=np.nan
```

Out[13]:

	alcohol	malic_acid	alcalinity_of_ash	magnesium	flavanoids	proanthocyanins	hue
0	NaN	1.78	11.2	100.0	2.76	1.28	1.05
1	NaN	2.36	18.6	101.0	3.24	2.81	1.03
2	NaN	1.95	16.8	NaN	3.49	2.18	0.86
3	13.24	2.59	21.0	NaN	2.69	1.82	1.04
4	14.20	1.76	15.2	112.0	3.39	1.97	1.05

Fill in the null values (NaN) with the number 10 in the alcohol column and 100 in magnesium column.

In [18]:

```
values={'alcohol':10,'magnesium':100}
df.fillna(value=values, inplace=True)
```

In [19]:

```
df.head()
```

Out[19]:

	alcohol	malic_acid	alcalinity_of_ash	magnesium	flavanoids	proanthocyanins	hue
0	10.00	1.78	11.2	100.0	2.76	1.28	1.05
1	10.00	2.36	18.6	101.0	3.24	2.81	1.03
2	10.00	1.95	16.8	100.0	3.49	2.18	0.86
3	13.24	2.59	21.0	100.0	2.69	1.82	1.04
4	14.20	1.76	15.2	112.0	3.39	1.97	1.05

Count the number of missing values in the entire dataset.

In [20]:

```
df.isna().sum()
```

Out[20]:

```
alcohol              0
malic_acid           0
alcalinity_of_ash    0
magnesium            0
flavanoids           0
proanthocyanins      0
```

```
hue                  0
dtype: int64
```

```
In [ ]:
```