saheart-data-set-lab-assignment

March 2, 2023

[3]:
```
# import pandas
import pandas as pd
```

[5]:
```
#The Data Set contains retrospective sample of males in a heart-disease␣
 ↪high-risk region of South Africa.(SAheart.csv)
SAheart=pd.read_csv("F:\KRISHNA\PG Data Engineering\Assignments\SAheart.csv")
SAheart
```

[5]:

	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
0	160	12.00	5.73	23.11	Present	49	25.30	97.20	52	Si
1	144	0.01	4.41	28.61	Absent	55	28.87	2.06	63	Si
2	118	0.08	3.48	32.28	Present	52	29.14	3.81	46	No
3	170	7.50	6.41	38.03	Present	51	31.99	24.26	58	Si
4	134	13.60	3.50	27.78	Present	60	25.99	57.34	49	Si
..	…	…	…	…	…	…	…	…	..	
457	214	0.40	5.98	31.72	Absent	64	28.45	0.00	58	No
458	182	4.20	4.41	32.10	Absent	52	28.61	18.72	52	Si
459	108	3.00	1.59	15.23	Absent	40	20.09	26.64	55	No
460	118	5.40	11.61	30.79	Absent	64	27.35	23.97	40	No
461	132	0.00	4.82	33.41	Present	62	14.70	0.00	46	Si

[462 rows x 10 columns]

[6]:
```
# On the basis of this, answer following:
# (1) How many records are in the dataset.

len(SAheart)
SAheart.shape
```

[6]:
```
(462, 10)
```

[7]:
```
# (2) Print metadata.
# data information
SAheart.info()

# data columns description
SAheart.columns
```

```
# describing columns
SAheart.describe()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 462 entries, 0 to 461
Data columns (total 10 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   sbp        462 non-null    int64
 1   tobacco    462 non-null    float64
 2   ldl        462 non-null    float64
 3   adiposity  462 non-null    float64
 4   famhist    462 non-null    object
 5   typea      462 non-null    int64
 6   obesity    462 non-null    float64
 7   alcohol    462 non-null    float64
 8   age        462 non-null    int64
 9   chd        462 non-null    object
dtypes: float64(5), int64(3), object(2)
memory usage: 36.2+ KB
```

[7]:

	sbp	tobacco	ldl	adiposity	typea	obesity	alcohol	age
count	462.000000	462.000000	462.000000	462.000000	462.000000	462.000000	462.000000	462.000000
mean	138.326840	3.635649	4.740325	25.406732	53.103896	26.044113	17.044394	42.816017
std	20.496317	4.593024	2.070909	7.780699	9.817534	4.213680	24.481059	14.608956
min	101.000000	0.000000	0.980000	6.740000	13.000000	14.700000	0.000000	15.000000
25%	124.000000	0.052500	3.282500	19.775000	47.000000	22.985000	0.510000	31.000000
50%	134.000000	2.000000	4.340000	26.115000	53.000000	25.805000	7.510000	45.000000
75%	148.000000	5.500000	5.790000	31.227500	60.000000	28.497500	23.892500	55.000000
max	218.000000	31.200000	15.330000	42.490000	78.000000	46.580000	147.190000	64.000000

[8]:
```
SAheart.describe(include=['O'])
```

[8]:

	famhist	chd
count	462	462
unique	2	2

```
top     Absent   No
freq       270  302
```

[9]:
```
import numpy as np
SAheart.quantile(np.arange(0,1,0.1))
# 20% of samples have alcohol on 0.0 and only 40% over 10.00
# 60% of samples are under 50 years old
```

[9]:

	sbp	tobacco	ldl	adiposity	typea	obesity	alcohol	age
0.0	101.0	0.000	0.980	6.740	13.0	14.700	0.000	15.0
0.1	118.0	0.000	2.510	13.713	41.0	21.142	0.000	18.0
0.2	122.0	0.000	3.104	17.930	46.0	22.320	0.000	28.0
0.3	126.0	0.400	3.543	21.145	49.0	23.633	1.385	34.0
0.4	130.0	1.024	3.950	23.874	51.0	24.804	3.248	40.0
0.5	134.0	2.000	4.340	26.115	53.0	25.805	7.510	45.0
0.6	138.0	3.436	4.890	28.082	56.0	26.706	11.830	49.0
0.7	144.0	4.500	5.457	30.057	58.0	27.807	19.298	54.0
0.8	154.0	6.156	6.138	32.472	61.0	29.114	27.770	58.0
0.9	166.0	9.090	7.383	35.352	65.0	30.965	47.510	61.0

[10]:
```
SAheart['chd'].value_counts()
```

[10]:
```
No    302
Si    160
Name: chd, dtype: int64
```

[11]:
```
SAheart['famhist'].value_counts()
```

[11]:
```
Absent     270
Present    192
Name: famhist, dtype: int64
```

[12]:
```
#206 samples have chd in their family history, maybe correlation features.
pd.crosstab(index = SAheart['famhist'], columns = SAheart["chd"])
```

[12]:

chd	No	Si
famhist		
Absent	206	64
Present	96	96

[13]:
```
# (4) Find Number of chd(Response,coronary heart disease) cases in different␣
 ↪age categories. Draw a barpot for that
chd_no = SAheart['chd'].value_counts()[0]
chd_no
```

[13]:
```
302
```

[14]:
```
chd_yes = SAheart['chd'].value_counts()[1]
chd_yes
```

[14]:
```
160
```

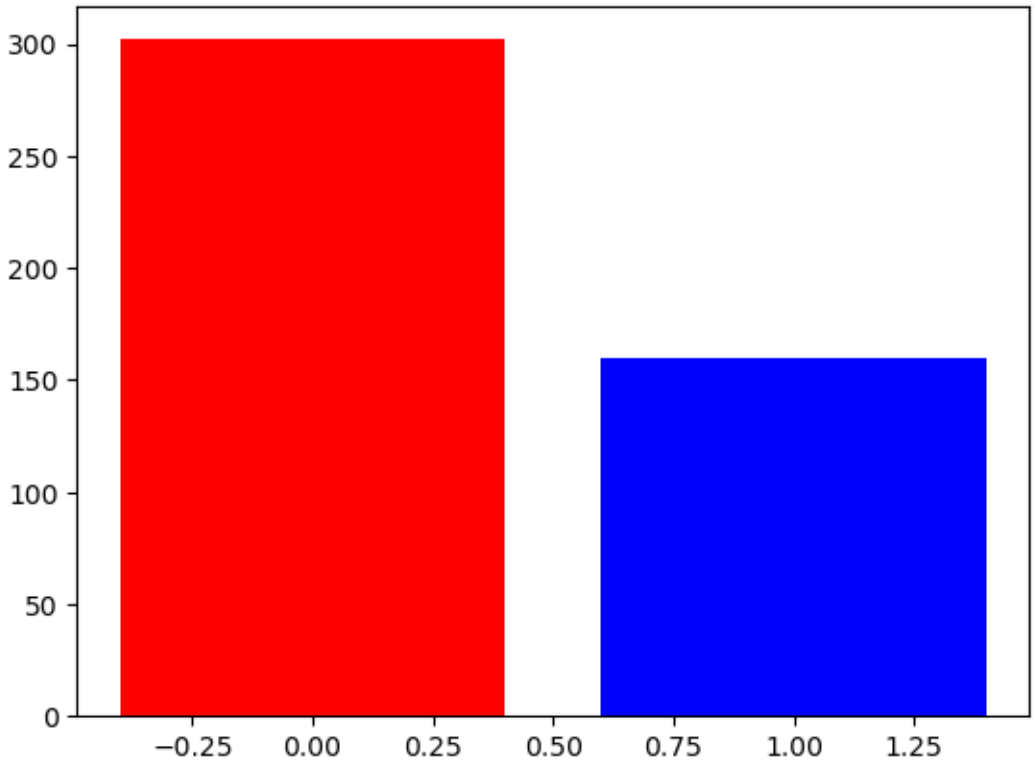
[29]:
```
np.arange(2)
```

[29]:
```
array([0, 1])
```

[16]:
```
import matplotlib.pyplot as plt

plt.bar([0,1], [chd_no, chd_yes], color = ["red", "blue"])
```

[16]:
```
<BarContainer object of 2 artists>
```

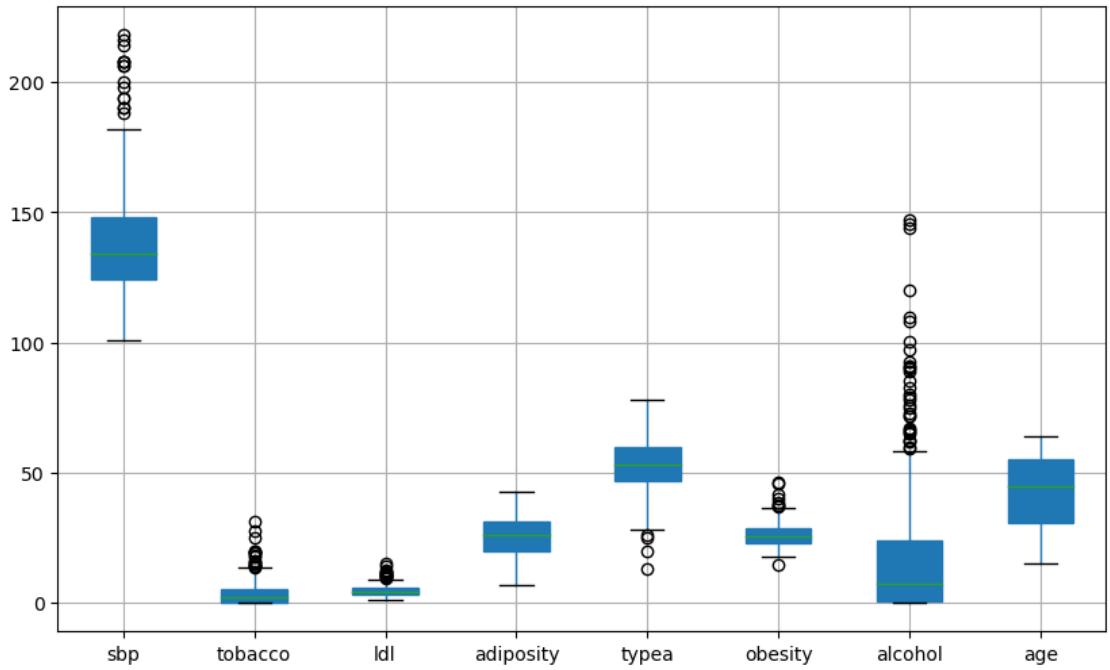


[17]:
```
plt.xticks([0,1], ["No","Si"])
```

[17]:
```
([<matplotlib.axis.XTick at 0x29d9ae35be0>,
  <matplotlib.axis.XTick at 0x29d9ae35bb0>],
 [Text(0, 0, 'No'), Text(1, 0, 'Si')])
```

[18]:
```
# Draw a boxplot to compare distributions of  ldl for different agegroups.
bplot2=SAheart.boxplot(figsize=(10,6),patch_artist=True)
bplot2
```

[18]: <AxesSubplot:>



[19]:
```
# (3) Add a new Colum called agegrp from age column as follows: (0-15) : young␣
 ↪(15-35):adults, (35-55):mid, (55-)old
agegroups = [] # make a new empty list for all the age groups

i = 0
for i in range(len(SAheart)):
    if SAheart['age'][i] <= 15:
        agegroups.append('young')
    elif SAheart['age'][i] <= 35:
        agegroups.append('adults')
    elif SAheart['age'][i] <= 55:
        agegroups.append('mid')
    else:
        agegroups.append('old')

SAheart['agegroup'] = agegroups # create a new column and define it as the␣
 ↪array we created and appended above
SAheart.head()
```

[19]:

	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd	\
0	160	12.00	5.73	23.11	Present	49	25.30	97.20	52	Si	
1	144	0.01	4.41	28.61	Absent	55	28.87	2.06	63	Si	
2	118	0.08	3.48	32.28	Present	52	29.14	3.81	46	No	
3	170	7.50	6.41	38.03	Present	51	31.99	24.26	58	Si	
4	134	13.60	3.50	27.78	Present	60	25.99	57.34	49	Si	

```
  agegroup
0      mid
1      old
2      mid
3      old
4      mid
```

[20]:
```
# (5) Draw a boxplot to compare distributions of  ldl for different agegroups.
SAheart.boxplot(by ='agegroup', column =['ldl'], grid = False)
```

[20]:
```
<AxesSubplot:title={'center':'ldl'}, xlabel='agegroup'>
```