```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import random as rd
```

```
df = pd.read_csv("/data.csv",encoding = "ISO-8859-1")
```

```
<ipython-input-2-820e9fd356b3>:1: DtypeWarning: Columns (0) have mixed types. Specify dtype option on import or set low_memo
  df = pd.read_csv("/data.csv",encoding = "ISO-8859-1")
```

```
dh = pd.read_csv("/heart.csv")
```

```
df
```

	stn_code	sampling_date	state	location	agency	type	so2	r
0	150.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	4.8	1
1	151.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	3.1	
2	152.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.2	2
3	150.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.3	1
4	151.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.7	
...	...	...	...	...	...	...	...	
435737	SAMP	24-12-15	West Bengal	ULUBERIA	West Bengal State Pollution Control Board	RIRUO	22.0	5
435738	SAMP	29-12-15	West Bengal	ULUBERIA	West Bengal State Pollution Control Board	RIRUO	20.0	4
435739	NaN	NaN	andaman-and-nicobar-islands	NaN	NaN	NaN	NaN	N
435740	NaN	NaN	Lakshadweep	NaN	NaN	NaN	NaN	N
435741	NaN	NaN	Tripura	NaN	NaN	NaN	NaN	N

435742 rows × 13 columns

```
dh
```

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thal
0	63	1	3	145	233	1	0	150	0	2.3	0	0	
1	37	1	2	130	250	0	1	187	0	3.5	0	0	
2	41	0	1	130	204	0	0	172	0	1.4	2	0	
3	56	1	1	120	236	0	1	178	0	0.8	2	0	
4	57	0	0	120	354	0	1	163	1	0.6	2	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
298	57	0	0	140	241	0	1	123	1	0.2	1	0	
299	45	1	3	110	264	0	1	132	0	1.2	1	0	
300	68	1	0	144	193	1	1	141	0	3.4	1	2	
301	57	1	0	130	131	0	1	115	1	1.2	1	1	
302	57	0	1	130	236	0	0	174	0	0.0	1	1	

Next steps: View recommended plots

```
df.head()
```

	stn_code	sampling_date	state	location	agency	type	so2	no2	rspm	s
0	150.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	4.8	17.4	NaN	N
1	151.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	3.1	7.0	NaN	N
2	152.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.2	28.5	NaN	N
3	150.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.3	14.7	NaN	N
4	151.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.7	7.5	NaN	N

```
dh.head()
```

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2

Next steps: View recommended plots

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 435742 entries, 0 to 435741
Data columns (total 13 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   stn_code                    291665 non-null  object
 1   sampling_date               435739 non-null  object
 2   state                       435742 non-null  object
 3   location                    435739 non-null  object
 4   agency                      286261 non-null  object
 5   type                        430349 non-null  object
 6   so2                         401096 non-null  float64
 7   no2                         419509 non-null  float64
```

```
 8   rspm                        395520 non-null  float64
 9   spm                         198355 non-null  float64
 10  location_monitoring_station  408251 non-null  object
 11  pm2_5                       9314 non-null    float64
 12  date                        435735 non-null  object
dtypes: float64(5), object(8)
memory usage: 43.2+ MB
```

```
dh.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    int64
 1   sex       303 non-null    int64
 2   cp        303 non-null    int64
 3   trtbps    303 non-null    int64
 4   chol      303 non-null    int64
 5   fbs       303 non-null    int64
 6   restecg   303 non-null    int64
 7   thalachh  303 non-null    int64
 8   exng      303 non-null    int64
 9   oldpeak   303 non-null    float64
 10  slp       303 non-null    int64
 11  caa       303 non-null    int64
 12  thall     303 non-null    int64
 13  output    303 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

```
df.isnull().sum()
```

```
stn_code                      144077
sampling_date                      3
state                              0
location                           3
agency                        149481
type                            5393
so2                            34646
no2                            16233
rspm                           40222
spm                           237387
location_monitoring_station    27491
pm2_5                         426428
date                               7
dtype: int64
```

```
dh.isnull().sum()
```

```
age         0
sex         0
cp          0
trtbps      0
chol        0
fbs         0
restecg     0
thalachh    0
exng        0
oldpeak     0
slp         0
caa         0
thall       0
output      0
dtype: int64
```

```
df.dropna()
```

```
dh.dropna()
```

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thal
0	63	1	3	145	233	1	0	150	0	2.3	0	0	
1	37	1	2	130	250	0	1	187	0	3.5	0	0	
2	41	0	1	130	204	0	0	172	0	1.4	2	0	
3	56	1	1	120	236	0	1	178	0	0.8	2	0	
4	57	0	0	120	354	0	1	163	1	0.6	2	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
298	57	0	0	140	241	0	1	123	1	0.2	1	0	
299	45	1	3	110	264	0	1	132	0	1.2	1	0	
300	68	1	0	144	193	1	1	141	0	3.4	1	2	
301	57	1	0	130	131	0	1	115	1	1.2	1	1	
302	57	0	1	130	236	0	0	174	0	0.0	1	1	

```
df1 = df.loc[111:999,['state','location','so2','rspm']]
df1
```

	state	location	so2	rspm
111	Andhra Pradesh	Hyderabad	4.9	NaN
112	Andhra Pradesh	Vishakhapatnam	NaN	NaN
113	Andhra Pradesh	Vishakhapatnam	11.2	NaN
114	Andhra Pradesh	Vishakhapatnam	4.5	NaN
115	Andhra Pradesh	Hyderabad	6.2	NaN
...	...	...	...	...
995	Andhra Pradesh	Hyderabad	2.8	NaN
996	Andhra Pradesh	Hyderabad	5.0	NaN
997	Andhra Pradesh	Hyderabad	5.5	NaN
998	Andhra Pradesh	Hyderabad	5.8	NaN
999	Andhra Pradesh	Hyderabad	5.9	NaN

889 rows × 4 columns

Next steps: View recommended plots

```
df2 = df.iloc[[1,3,5,4,22,43,54,67,7,8,9,50,10,11]]
df2
```

	stn_code	sampling_date	state	location	agency	type	so2	no2	rspm	spm	location_monitoring_station	pm2
1	151.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	3.1	7.0	NaN	NaN	NaN	N
3	150.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.3	14.7	NaN	NaN	NaN	N
5	152.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.4	25.7	NaN	NaN	NaN	N
4	151.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.7	7.5	NaN	NaN	NaN	N
22	152.0	September - M091990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	8.1	17.8	NaN	167.0	NaN	N
43	152.0	May - M051991	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	12.3	38.6	NaN	219.0	NaN	N
54	151.0	September - M091991	Andhra Pradesh	Hyderabad	NaN	Industrial Area	13.3	11.9	NaN	56.0	NaN	N
67	203.0	January - M011992	Andhra Pradesh	Hyderabad	Andhra Pradesh Pollution Control Board	NaN	35.8	12.5	NaN	261.0	NaN	N
7	151.0	April - M041990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.7	8.7	NaN	NaN	NaN	N
8	152.0	April - M041990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	4.2	23.0	NaN	NaN	NaN	N
9	151.0	May - M051990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.0	8.9	NaN	NaN	NaN	N
50	150.0	August - M081991	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	8.5	12.5	NaN	119.0	NaN	N
10	152.0	May - M051990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	3.6	18.6	NaN	NaN	NaN	N
11	150.0	June - M061990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	3.9	14.1	NaN	133.0	NaN	N

```
df_integration = pd.concat([df1,df2])
df_integration
```

	state	location	so2	rspm	stn_code	sampling_date	agency	type	no2	spm	location_monitoring_station
111	Andhra Pradesh	Hyderabad	4.9	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
112	Andhra Pradesh	Vishakhapatnam	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
113	Andhra Pradesh	Vishakhapatnam	11.2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
114	Andhra Pradesh	Vishakhapatnam	4.5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
115	Andhra Pradesh	Hyderabad	6.2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...
8	Andhra Pradesh	Hyderabad	4.2	NaN	152.0	April - M041990	NaN	Residential, Rural and other Areas	23.0	NaN	NaN
9	Andhra Pradesh	Hyderabad	4.0	NaN	151.0	May - M051990	NaN	Industrial Area	8.9	NaN	NaN
50	Andhra Pradesh	Hyderabad	8.5	NaN	150.0	August - M081991	NaN	Residential, Rural and other Areas	12.5	119.0	NaN
10	Andhra Pradesh	Hyderabad	3.6	NaN	152.0	May - M051990	NaN	Residential, Rural and other Areas	18.6	NaN	NaN
11	Andhra Pradesh	Hyderabad	3.9	NaN	150.0	June - M061990	NaN	Residential, Rural and other Areas	14.1	133.0	NaN

903 rows × 13 columns

```
df_integration.transpose()
```

	111	112	113	114	115	116	117	1:
state	Andhra Pradesh	Andhra Pradesh	Andhra Pradesh	Andhra Pradesh	Andhra Pradesh	Andhra Pradesh	Andhra Pradesh	Andh Prade:
location	Hyderabad	Vishakhapatnam	Vishakhapatnam	Vishakhapatnam	Hyderabad	Hyderabad	Hyderabad	Hyderaba
so2	4.9	NaN	11.2	4.5	6.2	7.3	7.3	13
rspm	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
stn_code	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
sampling_date	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
agency	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
type	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
no2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
spm	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
location_monitoring_station	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
pm2_5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
date	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na

13 rows × 903 columns


```
df.drop(columns = 'so2')
```

	stn_code	sampling_date	state	location	agency	type	no2	rspm	spm	location_monitoring_station
0	150.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	17.4	NaN	NaN	NaN
1	151.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	7.0	NaN	NaN	NaN
2	152.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	28.5	NaN	NaN	NaN
3	150.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	14.7	NaN	NaN	NaN
4	151.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	7.5	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...
435737	SAMP	24-12-15	West Bengal	ULUBERIA	West Bengal State Pollution Control Board	RIRUO	50.0	143.0	NaN	Inside Rampal Industries,ULUBERIA
435738	SAMP	29-12-15	West Bengal	ULUBERIA	West Bengal State Pollution Control Board	RIRUO	46.0	171.0	NaN	Inside Rampal Industries,ULUBERIA
435739	NaN	NaN	andaman-and-nicobar-islands	NaN	NaN	NaN	NaN	NaN	NaN	NaN
435740	NaN	NaN	Lakshadweep	NaN	NaN	NaN	NaN	NaN	NaN	NaN
435741	NaN	NaN	Tripura	NaN	NaN	NaN	NaN	NaN	NaN	NaN

435742 rows × 12 columns

```
df2.drop(1)
```

	stn_code	sampling_date	state	location	agency	type	so2	no2	rspm
3	150.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.3	14.7	NaN
5	152.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.4	25.7	NaN
4	151.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.7	7.5	NaN
22	152.0	September - M091990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	8.1	17.8	NaN
43	152.0	May - M051991	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	12.3	38.6	NaN
54	151.0	September - M091991	Andhra Pradesh	Hyderabad	NaN	Industrial Area	13.3	11.9	NaN
67	203.0	January - M011992	Andhra Pradesh	Hyderabad	Andhra Pradesh Pollution Control Board	NaN	35.8	12.5	NaN
7	151.0	April - M041990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.7	8.7	NaN
8	152.0	April - M041990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	4.2	23.0	NaN
9	151.0	May - M051990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.0	8.9	NaN
50	150.0	August - M081991	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	8.5	12.5	NaN
10	152.0	May - M051990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	3.6	18.6	NaN
11	150.0	June - M061990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	3.9	14.1	NaN

```
df.melt()
```

	variable	value
0	stn_code	150.0
1	stn_code	151.0
2	stn_code	152.0
3	stn_code	150.0
4	stn_code	151.0
...	...	...
5664641	date	2015-12-24
5664642	date	2015-12-29
5664643	date	NaN
5664644	date	NaN
5664645	date	NaN

5664646 rows × 2 columns

```
df_merged = pd.concat([df,dh])
```

```
df_merged
```

	stn_code	sampling_date	state	location	agency	type	so2	no2	rspm
0	150.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	4.8	17.4	NaN
1	151.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	3.1	7.0	NaN
2	152.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.2	28.5	NaN
3	150.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.3	14.7	NaN
4	151.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.7	7.5	NaN
...	...	...	...	...	...	...	...	...	...
298	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
299	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
300	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
301	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
302	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

436045 rows × 27 columns

```
dh['chol'].unique()
```

```
array([233, 250, 204, 236, 354, 192, 294, 263, 199, 168, 239, 275, 266,
       211, 283, 219, 340, 226, 247, 234, 243, 302, 212, 175, 417, 197,
       198, 177, 273, 213, 304, 232, 269, 360, 308, 245, 208, 264, 321,
       325, 235, 257, 216, 256, 231, 141, 252, 201, 222, 260, 182, 303,
       265, 309, 186, 203, 183, 220, 209, 258, 227, 261, 221, 205, 240,
       318, 298, 564, 277, 214, 248, 255, 207, 223, 288, 160, 394, 315,
       246, 244, 270, 195, 196, 254, 126, 313, 262, 215, 193, 271, 268,
       267, 210, 295, 306, 178, 242, 180, 228, 149, 278, 253, 342, 157,
       286, 229, 284, 224, 206, 167, 230, 335, 276, 353, 225, 330, 290,
       172, 305, 188, 282, 185, 326, 274, 164, 307, 249, 341, 407, 217,
       174, 281, 289, 322, 299, 300, 293, 184, 409, 259, 200, 327, 237,
       218, 319, 166, 311, 169, 187, 176, 241, 131])
```

```
dh['chol'].value_counts()
```

```
chol
204    6
197    6
234    6
269    5
254    5
      ..
284    1
224    1
167    1
276    1
131    1
Name: count, Length: 152, dtype: int64
```

```
dh['caa'].unique()
```

```
array([0, 2, 1, 3, 4])
```

```
dh['caa'].value_counts()
```

```
caa
0    175
1     65
2     38
3     20
4      5
Name: count, dtype: int64
```

```
from sklearn import linear_model, metrics
```

```
X = dh[["age"]]
Y = dh[["thall"]]
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X,Y,test_size=0.2,random_state=1)
```

```
len(X_train)
```

```
242
```

```
len(X_test)
```

```
61
```

```
dh.shape
```

```
(303, 14)
```

```
df.shape
```

```
(435742, 13)
```

```
reg = linear_model.LinearRegression()
```

```
print(X_train)
```

```
           age
62          52
127         67
111         57
287         57
108         50
..         ...
203         68
255         45
```