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

```
ds=pd.read_csv("/content/India Air Quality Data - India Air Quality Data.csv",encoding="ISO=8859-1")
```

```
<ipython-input-6-e14b9dc9ef77>:1: DtypeWarning: Columns (0) have mixed types. Specify dtype option on import or set low_m
  ds=pd.read_csv("/content/India Air Quality Data - India Air Quality Data.csv",encoding="ISO=8859-1")
```

```
ds
```

	stn_code	sampling_date	state	location	agency	type	so2
0	150.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	4.8
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
3	150.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.3
4	151.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.7
...	...	...	...	...	...	...	...
265418	733.0	06-05-14	Mizoram	Kolasib	Mizoram State Pollution Control Board	Residential, Rural and other Areas	2.0
265419	733.0	06-10-14	Mizoram	Kolasib	Mizoram State Pollution Control Board	Residential, Rural and other Areas	2.0
265420	733.0	06-12-14	Mizoram	Kolasib	Mizoram State Pollution Control Board	Residential, Rural and other Areas	2.0
265421	733.0	17-06-14	Mizoram	Kolasib	Mizoram State Pollution Control Board	Residential, Rural and other Areas	2.0
265422	733.0	19-06-14	Mizoram	Kolasib	Mizora	NaN	NaN

265423 rows × 13 columns

```
df=pd.read_csv("/content/heart - heart.csv")
```

```
df
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slo
0	52	1	0	125	212	0	1	168	0	1.0	
1	53	1	0	140	203	1	0	155	1	3.1	
2	70	1	0	145	174	0	1	125	1	2.6	
3	61	1	0	148	203	0	1	161	0	0.0	
4	62	0	0	138	294	1	1	106	0	1.9	
...	...	...	...	...	...	...	...	...	...	...	
1020	59	1	1	140	221	0	1	164	1	0.0	
1021	60	1	0	125	258	0	0	141	1	2.8	
1022	47	1	0	110	275	0	0	118	1	1.0	
1023	50	0	0	110	254	0	0	159	0	0.0	
1024	54	1	0	120	188	0	1	113	0	1.4	

1025 rows × 14 columns

Next steps: Generate code with df View recommended plots

```
ds.head()
```

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

```
df.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope
0	52	1	0	125	212	0	1	168	0	1.0	2
1	53	1	0	140	203	1	0	155	1	3.1	0
2	70	1	0	145	174	0	1	125	1	2.6	0
3	61	1	0	148	203	0	1	161	0	0.0	2
4	62	0	0	138	294	1	1	106	0	1.9	1

Next steps: Generate code with df View recommended plots

```
ds.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 265423 entries, 0 to 265422
Data columns (total 13 columns):
 #   Column                       Non-Null Count   Dtype  
---  ------                       --------------   -----  
 0   stn_code                     178784 non-null  object 
 1   sampling_date                265423 non-null  object 
 2   state                        265423 non-null  object 
 3   location                     265423 non-null  object 
 4   agency                       175537 non-null  object 
 5   type                         262329 non-null  object 
 6   so2                          241612 non-null  float64
 7   no2                          253860 non-null  float64
 8   rspm                         240712 non-null  float64
 9   spm                          119712 non-null  float64
 10  location_monitoring_station  248485 non-null  object 
 11  pm2_5                        5078 non-null    float64
 12  date                         265419 non-null  object 
dtypes: float64(5), object(8)
memory usage: 26.3+ MB
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1025 entries, 0 to 1024
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1025 non-null   int64  
 1   sex       1025 non-null   int64  
 2   cp        1025 non-null   int64  
 3   trestbps  1025 non-null   int64  
 4   chol      1025 non-null   int64  
 5   fbs       1025 non-null   int64  
 6   restecg   1025 non-null   int64  
 7   thalach   1025 non-null   int64  
 8   exang     1025 non-null   int64  
 9   oldpeak   1025 non-null   float64
 10  slope     1025 non-null   int64  
 11  ca        1025 non-null   int64  
 12  thal      1025 non-null   int64  
 13  target    1025 non-null   int64  
dtypes: float64(1), int64(13)
memory usage: 112.2 KB
```

```
ds.isnull().sum()
```

```
stn_code                        86639
sampling_date                       0
state                               0
location                            0
agency                          89886
type                             3094
so2                             23811
no2                             11563
rspm                            24711
spm                            145711
location_monitoring_station     16938
pm2_5                          260345
date                                4
dtype: int64
```

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

```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

```
ds.dropna()
```

```
   stn_code  sampling_date  state  location  agency  type  so2  no2  rspm  spm
```

```
df.dropna()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slo
0	52	1	0	125	212	0	1	168	0	1.0	
1	53	1	0	140	203	1	0	155	1	3.1	
2	70	1	0	145	174	0	1	125	1	2.6	
3	61	1	0	148	203	0	1	161	0	0.0	
4	62	0	0	138	294	1	1	106	0	1.9	
...	...	...	...	...	...	...	...	...	...	...	
1020	59	1	1	140	221	0	1	164	1	0.0	
1021	60	1	0	125	258	0	0	141	1	2.8	
1022	47	1	0	110	275	0	0	118	1	1.0	
1023	50	0	0	110	254	0	0	159	0	0.0	
1024	54	1	0	120	188	0	1	113	0	1.4	

1025 rows × 14 columns

```
ds1=ds.loc[111:999,['state', 'location', 'so2', 'rspm']]
```

```
ds2=ds.iloc[[1,3,5,4,22,43,54,67,7,8,9,50,10,11]]
```

```
ds1
```

	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

```
ds2
```

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

```
ds_integration=pd.concat([ds1,ds2])
```

```
ds_integration
```

	state	location	so2	rspm	stn_code	sampling_date	agency	
111	Andhra Pradesh	Hyderabad	4.9	NaN	NaN	NaN	NaN	
112	Andhra Pradesh	Vishakhapatnam	NaN	NaN	NaN	NaN	NaN	
113	Andhra Pradesh	Vishakhapatnam	11.2	NaN	NaN	NaN	NaN	
114	Andhra Pradesh	Vishakhapatnam	4.5	NaN	NaN	NaN	NaN	
115	Andhra Pradesh	Hyderabad	6.2	NaN	NaN	NaN	NaN	
...	...	...	...	...	...	...	...	
8	Andhra Pradesh	Hyderabad	4.2	NaN	152.0	April - M041990	NaN	Resi Ru othe
9	Andhra Pradesh	Hyderabad	4.0	NaN	151.0	May - M051990	NaN	In
50	Andhra Pradesh	Hyderabad	8.5	NaN	150.0	August - M081991	NaN	Resi Ru othe
10	Andhra Pradesh	Hyderabad	3.6	NaN	152.0	May - M051990	NaN	Resi Ru othe
11	Andhra Pradesh	Hyderabad	3.9	NaN	150.0	June - M061990	NaN	Resi Ru othe

903 rows × 13 columns

ds_integration.transpose()

```
ds.drop(columns = "so2")
```

	stn_code	sampling_date	state	location	agency	type	no2
0	150.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	17.4
1	151.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	7.0
2	152.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	28.5
3	150.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	14.7
4	151.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	7.5
...	...	...	...	...	...	...	...
265418	733.0	06-05-14	Mizoram	Kolasib	Mizoram State Pollution Control Board	Residential, Rural and other Areas	5.0
265419	733.0	06-10-14	Mizoram	Kolasib	Mizoram State Pollution Control Board	Residential, Rural and other Areas	5.0
265420	733.0	06-12-14	Mizoram	Kolasib	Mizoram State Pollution Control Board	Residential, Rural and other Areas	5.0
265421	733.0	17-06-14	Mizoram	Kolasib	Mizoram State Pollution Control Board	Residential, Rural and other Areas	5.0
265422	733.0	19-06-14	Mizoram	Kolasib	Mizora	NaN	NaN

265423 rows × 12 columns

```
ds2.drop(1)
```

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

```
ds.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
...	...	...
3450494	date	2014-05-06
3450495	date	2014-10-06
3450496	date	2014-12-06
3450497	date	2014-06-17
3450498	date	NaN

3450499 rows × 2 columns

```
ds_merged=pd.concat([ds,df])
```

```
ds_merged
```

	stn_code	sampling_date	state	location	agency	type	so2	n
0	150.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	4.8	17
1	151.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	3.1	7
2	152.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.2	28
3	150.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.3	14
4	151.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.7	7
...	...	...	...	...	...	...	...	
1020	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
1021	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
1022	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
1023	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
1024	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na

266448 rows × 27 columns

```
df['ca'].unique()
```

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

```
df.ca.value_counts()
```

```
0    578
1    226
2    134
3     69
4     18
Name: ca, dtype: int64
```

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

```
X=df[["age"]]
```

```
Y=df[["thal"]]
```

```
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)
```

```
820
```

```
len(X_test)
```

```
205
```

```
ds.shape
```

```
(265423, 13)
```

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

```
print(X_train)
```

```
     age
880   57
358   59
772   62
682   59
848   58
..   ...
905   64
767   68
72    56
908   62
235   64

[820 rows x 1 columns]
```

```
model=reg.fit(X_train,Y_train)
```

```
r_sq=reg.score(X_train,Y_train)
```

```
print("determination coefficient:",r_sq)
```

```
determination coefficient: 0.008792008347529245
```

```
print("intercept:",model.intercept_)
```

```
intercept: [1.97461867]
```

```
print("slope:",model.coef_)
```

```
slope: [[0.00633286]]
```

```
Y_pred=model.predict(X_test)
```

```
print('predicted response: ',Y_pred,sep='\n')
```

```
 [2.3609229 ]
 [2.32292576]
 [2.20893435]
 [2.39258718]
 [2.30392719]
 [2.34825718]
 [2.24693148]
 [2.33559147]
 [2.34192433]
 [2.31026005]
 [2.34192433]
 [2.34825718]
 [2.39892003]
 [2.34825718]
 [2.3165929 ]
 [2.39892003]
 [2.42425146]
 [2.32292576]
 [2.26593005]
 [2.25326434]
 [2.32925862]
 [2.39258718]
 [2.3165929 ]
 [2.31026005]
 [2.22160006]
 [2.20893435]
 [2.37992147]
 [2.23426577]
 [2.30392719]
 [2.35459004]
 [2.2152672 ]
 [2.19626863]
 [2.4179186 ]
 [2.34192433]
 [2.29759433]
```