```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn import linear_model
```

```
df = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/AirQualityUCI.csv')
df
```

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO
0	10/03/2004	18:00:00	2.6	1360.0	150.0	11.9	1046.0	166.0	1056.0	113.0	169
1	10/03/2004	19:00:00	2.0	1292.0	112.0	9.4	955.0	103.0	1174.0	92.0	155
2	10/03/2004	20:00:00	2.2	1402.0	88.0	9.0	939.0	131.0	1140.0	114.0	155
3	10/03/2004	21:00:00	2.2	1376.0	80.0	9.2	948.0	172.0	1092.0	122.0	158
4	10/03/2004	22:00:00	1.6	1272.0	51.0	6.5	836.0	131.0	1205.0	116.0	149
...	...	...	...	...	...	...	...	...	...	...	
9466	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
9467	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
9468	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
9469	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
9470	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N

9471 rows × 17 columns

```
df.dtypes
```

```
Date              object
Time              object
CO(GT)           float64
PT08.S1(CO)      float64
NMHC(GT)         float64
C6H6(GT)         float64
PT08.S2(NMHC)    float64
NOx(GT)          float64
PT08.S3(NOx)     float64
NO2(GT)          float64
PT08.S4(NO2)     float64
PT08.S5(O3)      float64
T                float64
RH               float64
AH               float64
Unnamed: 15      float64
Unnamed: 16      float64
dtype: object
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9471 entries, 0 to 9470
Data columns (total 17 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Date           9357 non-null   object 
 1   Time           9357 non-null   object 
 2   CO(GT)         9357 non-null   float64
 3   PT08.S1(CO)    9357 non-null   float64
 4   NMHC(GT)       9357 non-null   float64
 5   C6H6(GT)       9357 non-null   float64
 6   PT08.S2(NMHC)  9357 non-null   float64
 7   NOx(GT)        9357 non-null   float64
 8   PT08.S3(NOx)   9357 non-null   float64
 9   NO2(GT)        9357 non-null   float64
 10  PT08.S4(NO2)   9357 non-null   float64
 11  PT08.S5(O3)    9357 non-null   float64
 12  T              9357 non-null   float64
 13  RH             9357 non-null   float64
 14  AH             9357 non-null   float64
 15  Unnamed: 15    0 non-null      float64
 16  Unnamed: 16    0 non-null      float64
dtypes: float64(15), object(2)
memory usage: 1.2+ MB
```

```
#Simply drop the whole row as the missing values is less than 5% of the total data set
df.dropna(subset = ['AH'], inplace = True, axis = 0 )

#Resetting index, as we deleted some rows
df.reset_index(drop = True, inplace = True)
```

```
df.tail(10)
```

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO
9347	04/04/2005	05:00:00	0.5	888.0	-200.0	1.3	528.0	77.0	1077.0	53.0	98
9348	04/04/2005	06:00:00	1.1	1031.0	-200.0	4.4	730.0	182.0	760.0	93.0	112
9349	04/04/2005	07:00:00	4.0	1384.0	-200.0	17.4	1221.0	594.0	470.0	155.0	160
9350	04/04/2005	08:00:00	5.0	1446.0	-200.0	22.4	1362.0	586.0	415.0	174.0	177
9351	04/04/2005	09:00:00	3.9	1297.0	-200.0	13.6	1102.0	523.0	507.0	187.0	137
9352	04/04/2005	10:00:00	3.1	1314.0	-200.0	13.5	1101.0	472.0	539.0	190.0	137
9353	04/04/2005	11:00:00	2.4	1163.0	-200.0	11.4	1027.0	353.0	604.0	179.0	126
9354	04/04/2005	12:00:00	2.4	1142.0	-200.0	12.4	1063.0	293.0	603.0	175.0	124
9355	04/04/2005	13:00:00	2.1	1003.0	-200.0	9.5	961.0	235.0	702.0	156.0	104
9356	04/04/2005	14:00:00	2.2	1071.0	-200.0	11.9	1047.0	265.0	654.0	168.0	112

```
df.drop(['Unnamed: 15', 'Unnamed: 16'], axis = 1, inplace=True)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9357 entries, 0 to 9356
Data columns (total 15 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Date           9357 non-null   object 
 1   Time           9357 non-null   object 
 2   CO(GT)         9357 non-null   float64
 3   PT08.S1(CO)    9357 non-null   float64
 4   NMHC(GT)       9357 non-null   float64
 5   C6H6(GT)       9357 non-null   float64
 6   PT08.S2(NMHC)  9357 non-null   float64
 7   NOx(GT)        9357 non-null   float64
 8   PT08.S3(NOx)   9357 non-null   float64
 9   NO2(GT)        9357 non-null   float64
 10  PT08.S4(NO2)   9357 non-null   float64
 11  PT08.S5(O3)    9357 non-null   float64
 12  T              9357 non-null   float64
 13  RH             9357 non-null   float64
 14  AH             9357 non-null   float64
dtypes: float64(13), object(2)
memory usage: 1.1+ MB
```

```
df['Date'] = df['Date'].astype('category')
df['Date'] = df['Date'].cat.codes

df['Time'] = df['Time'].astype('category')
df['Time'] = df['Time'].cat.codes
```

```
df.tail(10)
```

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08
9347	43	5	0.5	888.0	-200.0	1.3	528.0	77.0	1077.0	53.0	987.0	
9348	43	6	1.1	1031.0	-200.0	4.4	730.0	182.0	760.0	93.0	1129.0	
9349	43	7	4.0	1384.0	-200.0	17.4	1221.0	594.0	470.0	155.0	1600.0	

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

```
Date             0
Time             0
CO(GT)           0
PT08.S1(CO)      0
NMHC(GT)         0
C6H6(GT)         0
PT08.S2(NMHC)    0
NOx(GT)          0
PT08.S3(NOx)     0
NO2(GT)          0
PT08.S4(NO2)     0
PT08.S5(O3)      0
T                0
RH               0
AH               0
dtype: int64
```

```
X = df.drop(columns = 'AH')
X
```

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08
0	114	18	2.6	1360.0	150.0	11.9	1046.0	166.0	1056.0	113.0	1692.0	
1	114	19	2.0	1292.0	112.0	9.4	955.0	103.0	1174.0	92.0	1559.0	
2	114	20	2.2	1402.0	88.0	9.0	939.0	131.0	1140.0	114.0	1555.0	
3	114	21	2.2	1376.0	80.0	9.2	948.0	172.0	1092.0	122.0	1584.0	
4	114	22	1.6	1272.0	51.0	6.5	836.0	131.0	1205.0	116.0	1490.0	
...	...	...	...	...	...	...	...	...	...	...	...	
9352	43	10	3.1	1314.0	-200.0	13.5	1101.0	472.0	539.0	190.0	1374.0	
9353	43	11	2.4	1163.0	-200.0	11.4	1027.0	353.0	604.0	179.0	1264.0	
9354	43	12	2.4	1142.0	-200.0	12.4	1063.0	293.0	603.0	175.0	1241.0	
9355	43	13	2.1	1003.0	-200.0	9.5	961.0	235.0	702.0	156.0	1041.0	
9356	43	14	2.2	1071.0	-200.0	11.9	1047.0	265.0	654.0	168.0	1129.0	

9357 rows × 14 columns

```
y = df['AH']
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
```

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X_train_scaler = scaler.fit_transform(X_train)
X_test_scaler = scaler.transform(X_test)
```

```
from sklearn.linear_model import LinearRegression
lr = LinearRegression()
```

```
# Fitting our training data to our linear regression model
lr.fit(X_train_scaler, y_train)
```

```
▾ LinearRegression
LinearRegression()
```

```
c = lr.intercept_
c
```

```
-6.953743563902885
```

```
m = lr.coef_
```

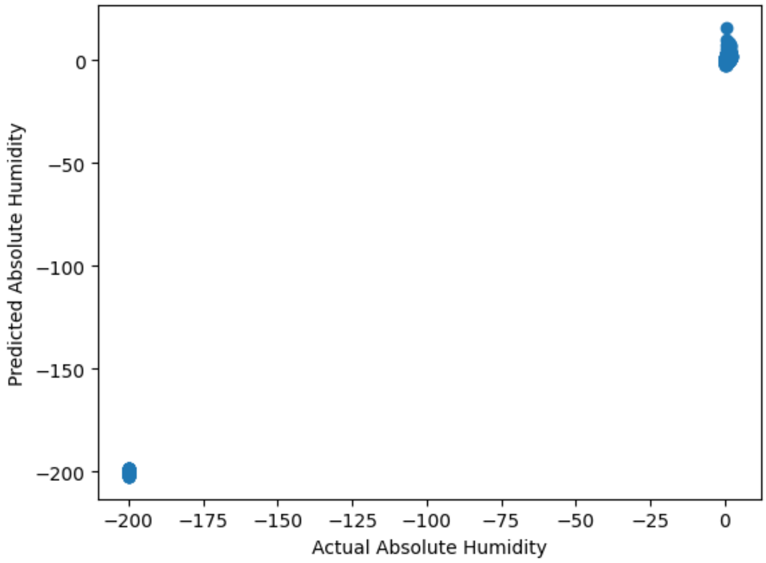
```
m
```

```
array([-5.21915711e-02,  1.52694065e-01, -5.57130177e-03, -2.76042444e-01,
        2.63557243e-01,  3.20174099e+01, -6.21895189e+00, -8.43389953e-01,
       -2.97894262e-01,  6.63583247e-01, -2.26458983e+00,  4.44067865e-01,
        9.86181865e+00,  3.95902365e+00])
```

```
y_pred_train = lr.predict(X_train_scaler)
y_pred_train
```

```
array([0.87446851, 2.32166635, 1.568725  , ..., 2.23828705, 1.74218483,
       1.86518636])
```

```
plt.scatter(y_train,y_pred_train)
plt.xlabel("Actual Absolute Humidity")
plt.ylabel("Predicted Absolute Humidity")
plt.show()
```

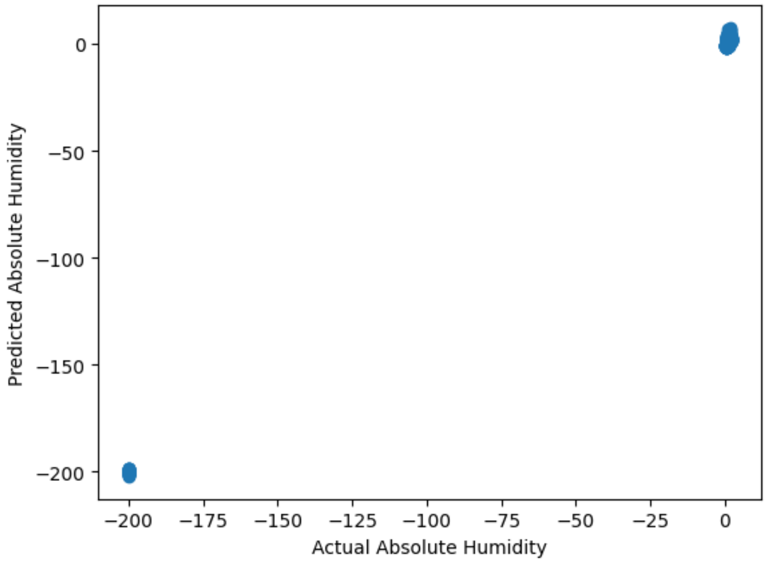


```
from sklearn.metrics import r2_score
r2_score(y_train, y_pred_train)
```

```
0.9993336584579668
```

```
y_pred_test = lr.predict(X_test_scaler)
```

```
plt.scatter(y_test,y_pred_test)
plt.xlabel("Actual Absolute Humidity")
plt.ylabel("Predicted Absolute Humidity")
plt.show()
```



```
r2_score(y_test, y_pred_test)
```

```
0.9993618893773653
```

✓ 0s completed at 19:24 ● ×