FirstStepsSols

October 21, 2022

[1]:
```
import numpy as np
import pandas as pd
```

[2]:
```
df = pd.read_csv('ozone.csv')
df
```

[2]:

	Ozone	Temp	InvHt	Pres	Vis	Hgt	Hum	InvTmp	Wind
0	3	40	2693	-25	250	5710	28	47.66	4
1	5	45	590	-24	100	5700	37	55.04	3
2	5	54	1450	25	60	5760	51	57.02	3
3	6	35	1568	15	60	5720	69	53.78	4
4	4	45	2631	-33	100	5790	19	54.14	6
..	…	…	…	…	…	…	…	…	…
325	8	50	2851	-5	70	5630	50	50.00	4
326	2	51	111	-14	200	5730	53	72.50	3
327	3	51	5000	-36	70	5690	23	51.26	3
328	5	50	3704	18	40	5650	61	46.94	3
329	1	39	5000	8	100	5550	85	39.92	4

[330 rows x 9 columns]

1 What is the range of each variable?

- What is the range of each input variable? What is the mean and standard deviation of each variable?

  useful functions: describe

[3]:
```
df.describe()
```

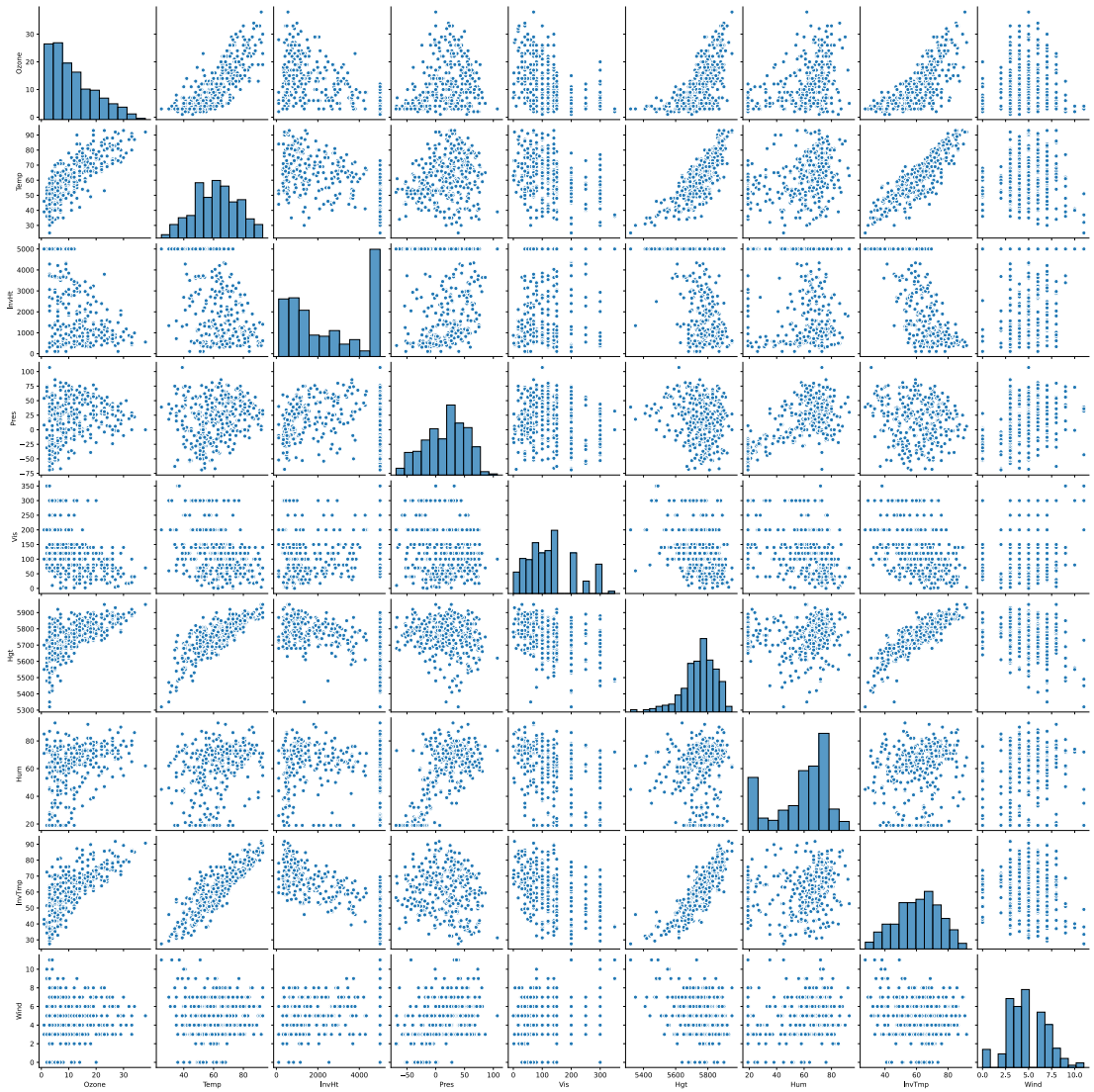
[3]:

	Ozone	Temp	InvHt	Pres	Vis	\
count	330.000000	330.000000	330.000000	330.000000	330.000000	
mean	11.775758	61.754545	2572.875758	17.369697	124.533333	
std	8.011277	14.458737	1803.885870	35.717181	79.362393	
min	1.000000	25.000000	111.000000	-69.000000	0.000000	
25%	5.000000	51.000000	877.500000	-9.000000	70.000000	
50%	10.000000	62.000000	2112.500000	24.000000	120.000000	
75%	17.000000	72.000000	5000.000000	44.750000	150.000000	

```
max      38.000000    93.000000  5000.000000  107.000000  350.000000

                Hgt         Hum      InvTmp         Wind
count    330.000000  330.000000  330.000000  330.000000
mean    5750.484848   58.130303   61.008909    4.848485
std      105.708241   19.865000   13.802296    2.116963
min     5320.000000   19.000000   27.500000    0.000000
25%     5690.000000   47.000000   51.260000    3.000000
50%     5760.000000   64.000000   62.150000    5.000000
75%     5830.000000   73.000000   70.520000    6.000000
max     5950.000000   93.000000   91.760000   11.000000
```

- Consult the visual plot of each pair of features below. Based on the corresponding scatterplot, what do you expect to be the correlation between Temperature and Height?



The above plot below is generated using the following code:

```
from seaborn import pairplot
from matplotlib import pyplot as plt
pairplot(df)
plt.savefig('pairplot.svg')
```

Answer: * By examining the corresponding plot, we see a very strong correlation, and in fact one close to a linear one.

- Compute the correlation matrix of each feature. What is the correlation between Temperature and Height? Does it match your expectations? *Useful commands:* pandas.DataFrame.corr

[4]:
```
df.corr()
```

[4]:

	Ozone	Temp	InvHt	Pres	Vis	Hgt	Hum	\
Ozone	1.000000	0.780703	-0.589534	0.214046	-0.440989	0.607344	0.449224	
Temp	0.780703	1.000000	-0.532645	0.189242	-0.387721	0.808059	0.340474	
InvHt	-0.589534	-0.532645	1.000000	0.037078	0.386686	-0.504835	-0.242328	
Pres	0.214046	0.189242	0.037078	1.000000	-0.125855	-0.148071	0.647789	
Vis	-0.440989	-0.387721	0.386686	-0.125855	1.000000	-0.360080	-0.401008	
Hgt	0.607344	0.808059	-0.504835	-0.148071	-0.360080	1.000000	0.074485	
Hum	0.449224	0.340474	-0.242328	0.647789	-0.401008	0.074485	1.000000	
InvTmp	0.745578	0.864787	-0.776933	-0.095060	-0.422372	0.852021	0.203648	
Wind	0.002471	-0.005886	0.196746	0.341951	0.127702	-0.225821	0.222868	

	InvTmp	Wind
Ozone	0.745578	0.002471
Temp	0.864787	-0.005886
InvHt	-0.776933	0.196746
Pres	-0.095060	0.341951
Vis	-0.422372	0.127702
Hgt	0.852021	-0.225821
Hum	0.203648	0.222868
InvTmp	1.000000	-0.159814
Wind	-0.159814	1.000000

Answer: * By examining the corresponding cell, we verify that the Pearson correlation between the features Temp and Hgt is 0.808. This rather high value is experimentally corroborating our intuition that arose from the visual inspection.