# 数据预处理和可视化作业-2

**1. 对 HUMI、PRES、TEMP 三列，进行线性插值处理。并对其中**

**超过 3 倍标准差的高度异常数据，修改为 3 倍标准差的数值。**

**2. 假设 PM 指数最高为 500，对 PM\_Dongsi、**

**PM\_Dongsihuan、 PM\_Nongzhanguan 三列中超过 500 的**

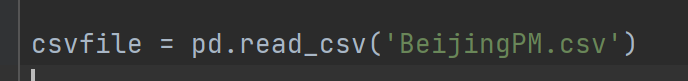
**数据，修改为 500PM 指数 进行异常值的处理。**

**3. 修改 cbwd 列中值为“cv”的单元格，其值用后项数据填充。**

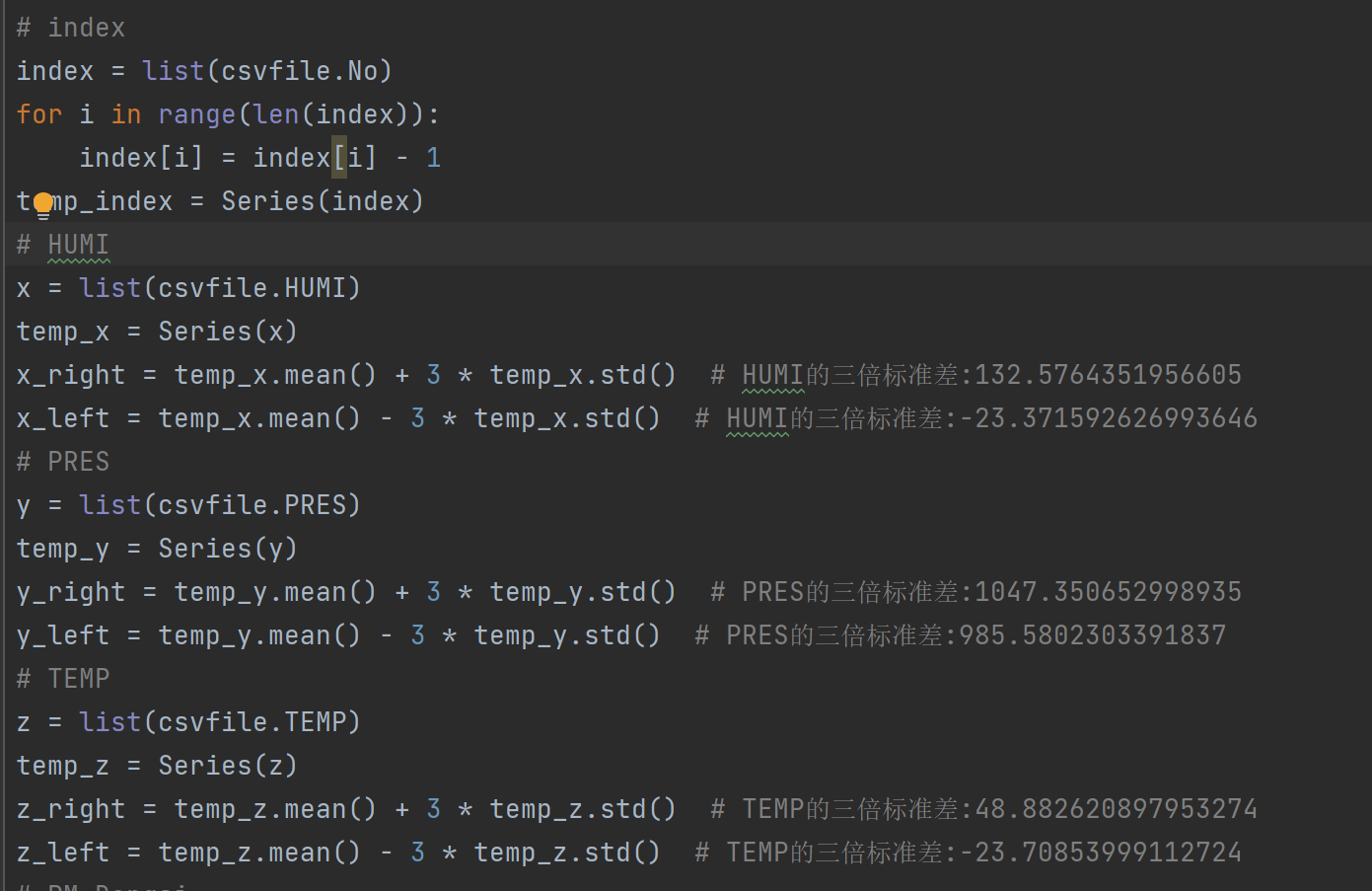
这三个问题的代码都写在了h\_p\_t.py里，下面对应每个问题，解释一下对应的代码：

首先是第一个问题：

第一步先打开csv文件，读入数据：



然后对所要处理的三列数据进行提取，同时也把No这一列也提取出来，以备后面使用。并且求出三列数据的三倍标准差：



由于这三列数据的处理类似，故设计了一个函数来进行处理数据，避免了代码重复。

该函数需要传入原始数据x，Series形式的原始数据temp\_x，以及三倍标准差x\_right和x\_left。

函数思路就是先将不是缺失值的原始数据读出来存到列表Y\_x中，

再将这些不是缺失值的原始数据的No栏读出来存到列表X\_index中，

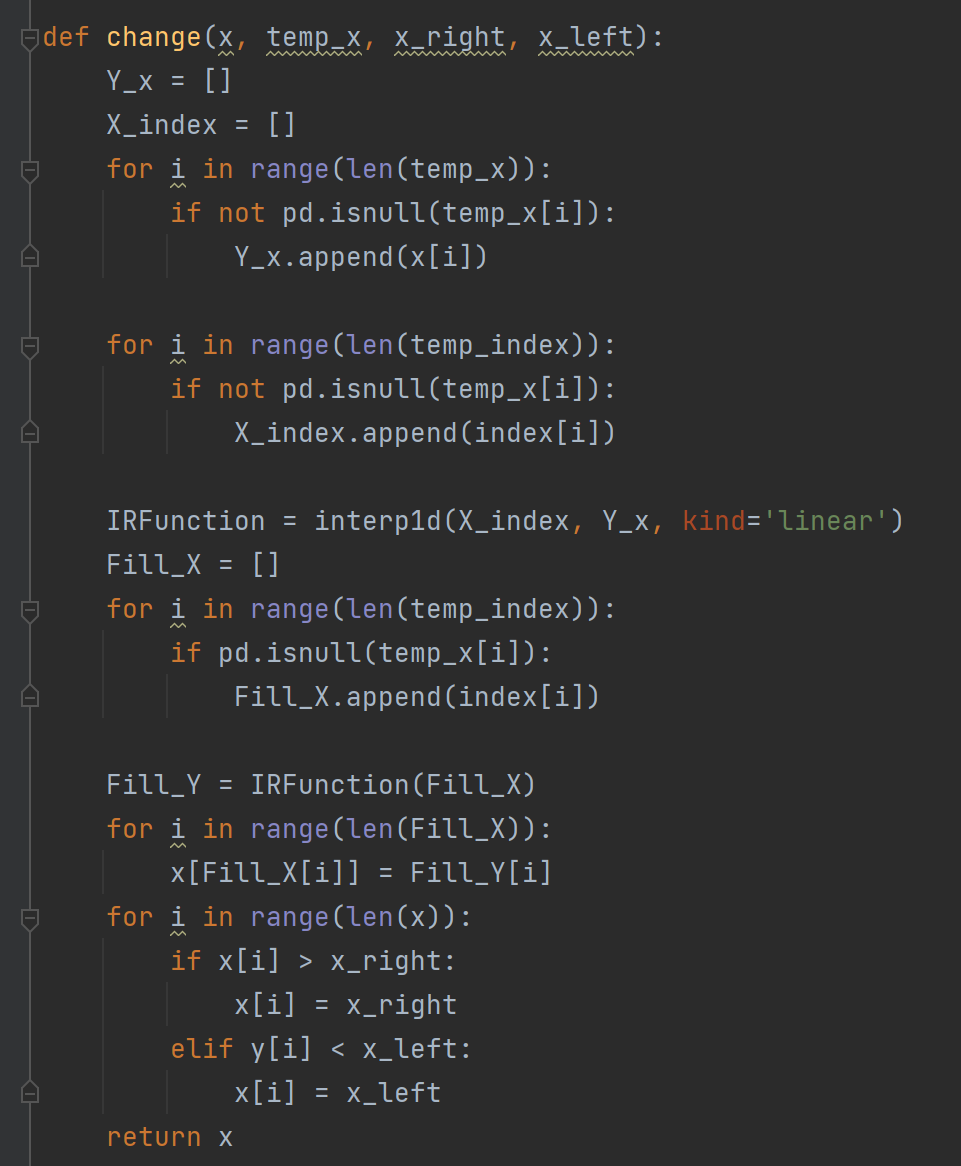
这两步是为了构造插值函数，为缺失值拟合数据。

之后构造插值函数。

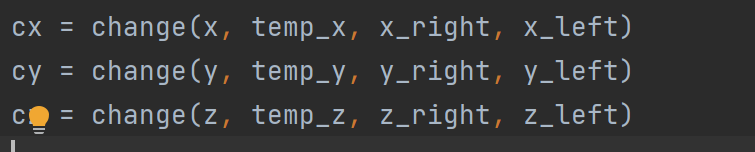
在之后获取原始数据中为缺失值的数据的No，存到列表里，

然后通过插值函数为这个下标数组求出拟合值数组，存到列表里。

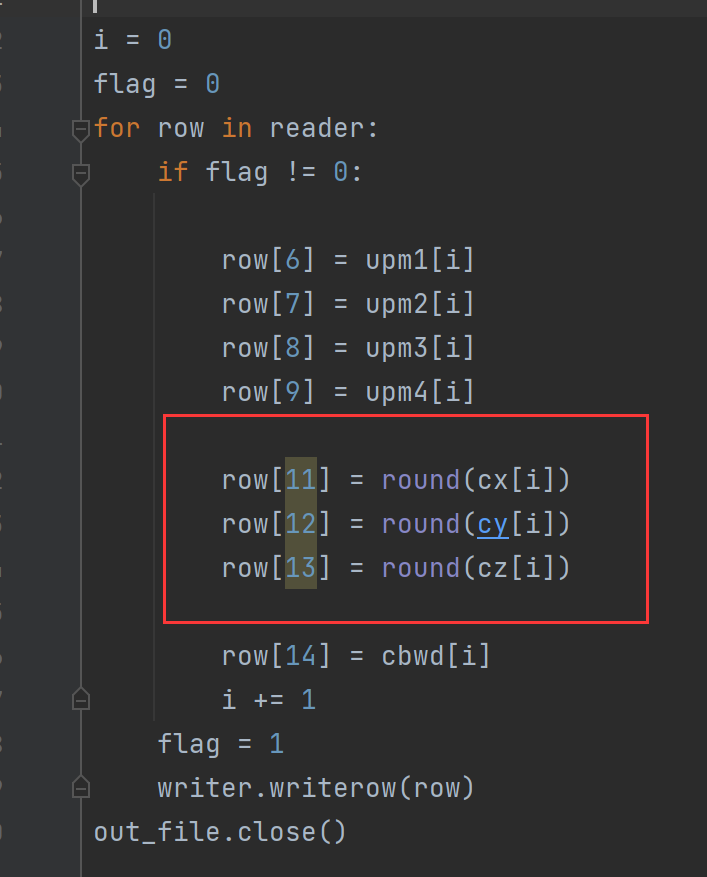
最后将拟合值覆盖到原始数据中，然后在判断是不是超过了三倍标准差，在决定要不要修改。最后返回修改后的数据。



为三列数据各调用一次上述函数，将返回值保存。



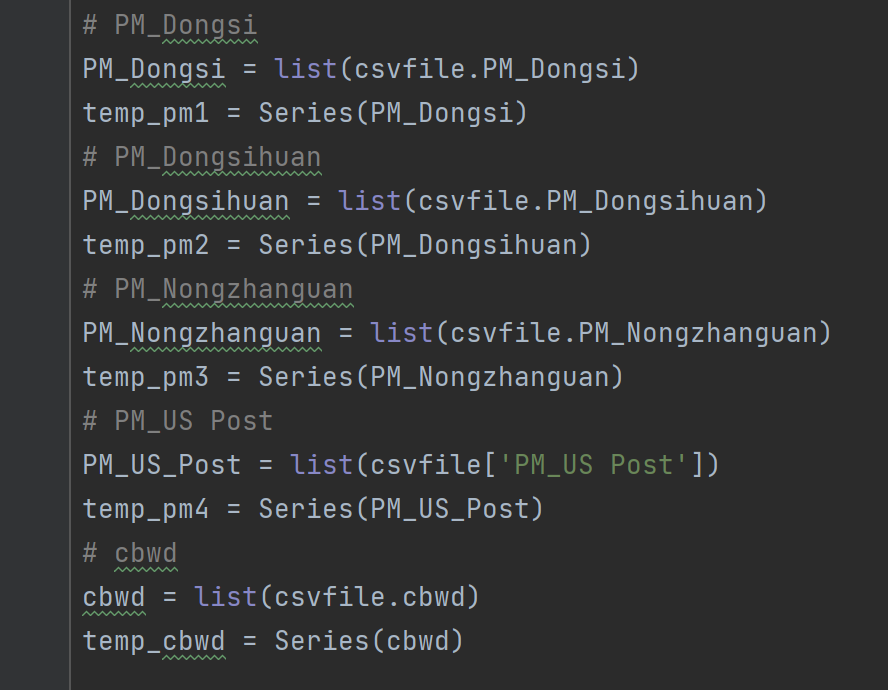
然后向新的csv文件写入：



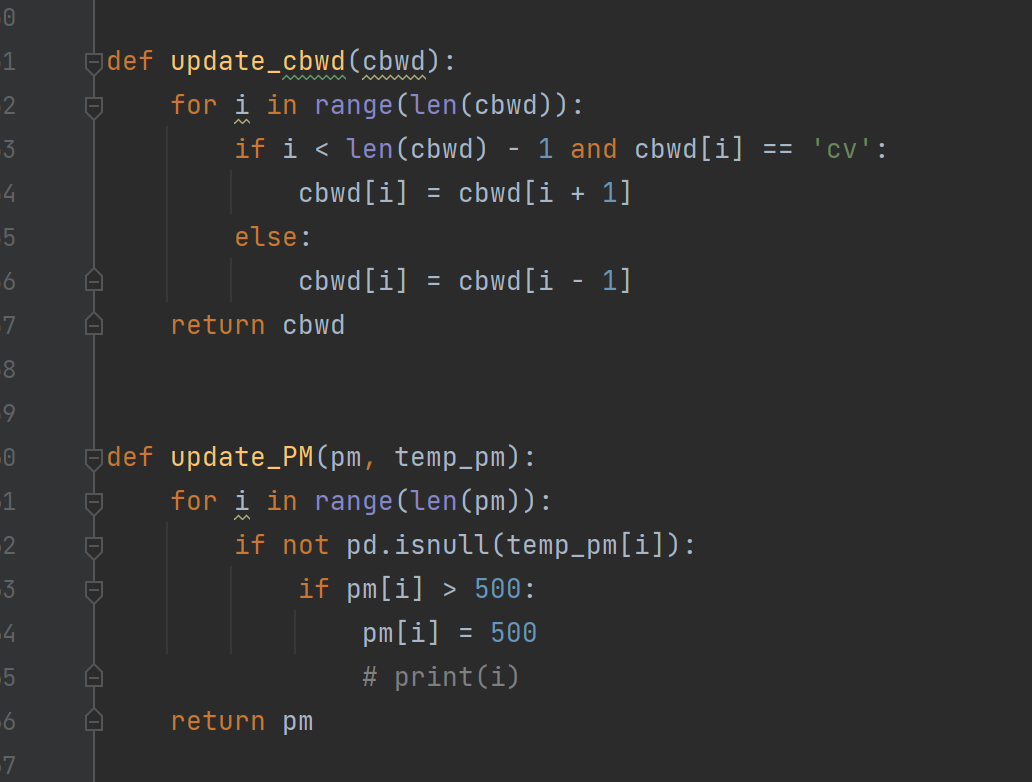
第一个问题到此结束。

第二第三个问题类似第一个：

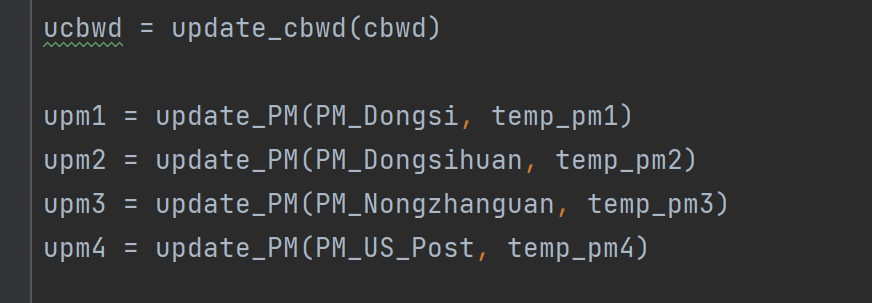
先获取原始数据：



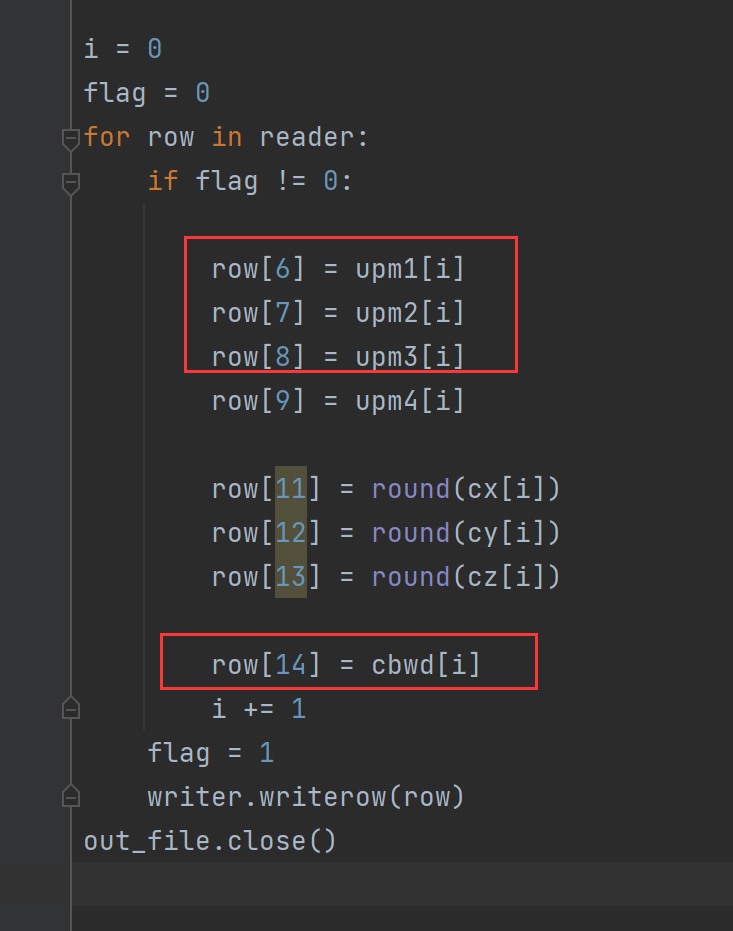
定义对数据进行操作的函数：



调用函数：



写入：



到此，前三个问题均解决。

**4. 对 DEWP 和 TEMP 两列，进行 0-1 归一化及 Z-Score 归一化处理。 结果使用散点图的形式表示（参考 PPT 第 19 页图形上**

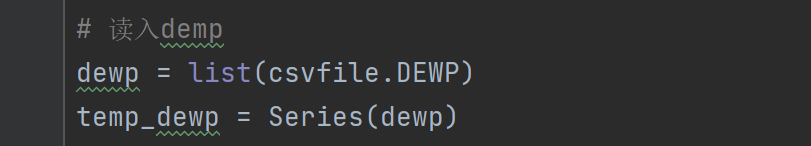
**半部分的表现形式）。**

解决第四个问题的代码在d\_t\_scaler.py里。

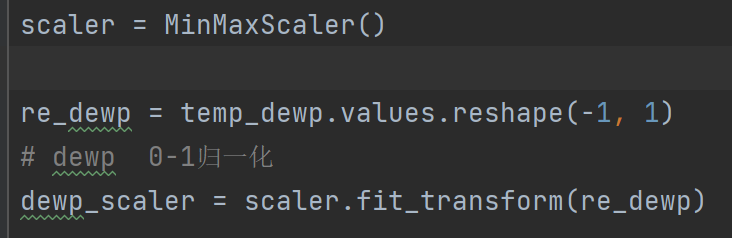
首先读入上述处理后的数据文件out.csv ：



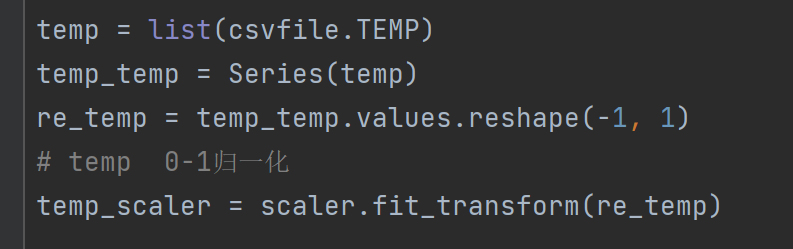
在获取要处理的数据：



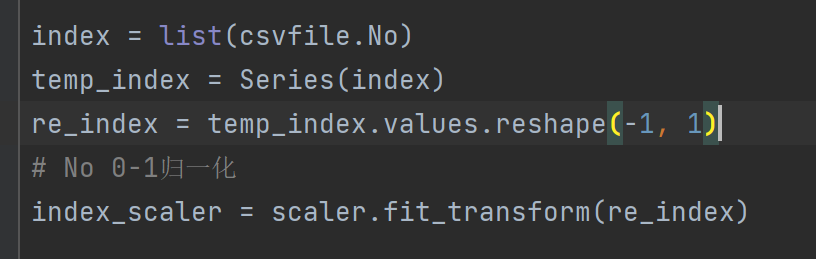
然后对该数据进行0-1归一化：



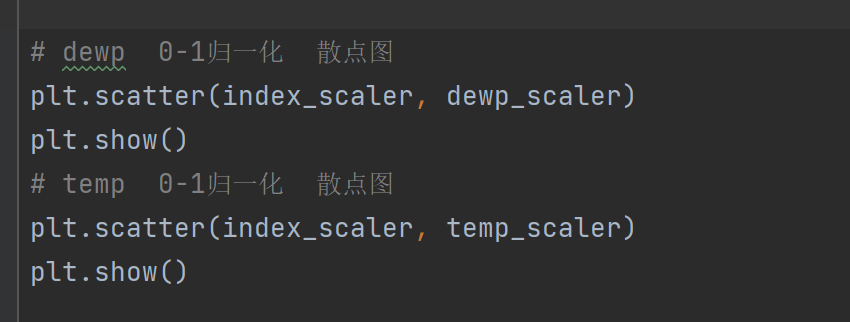
对另一列数据进行相同的操作：



同时也要对No列进行归一化，做X轴：

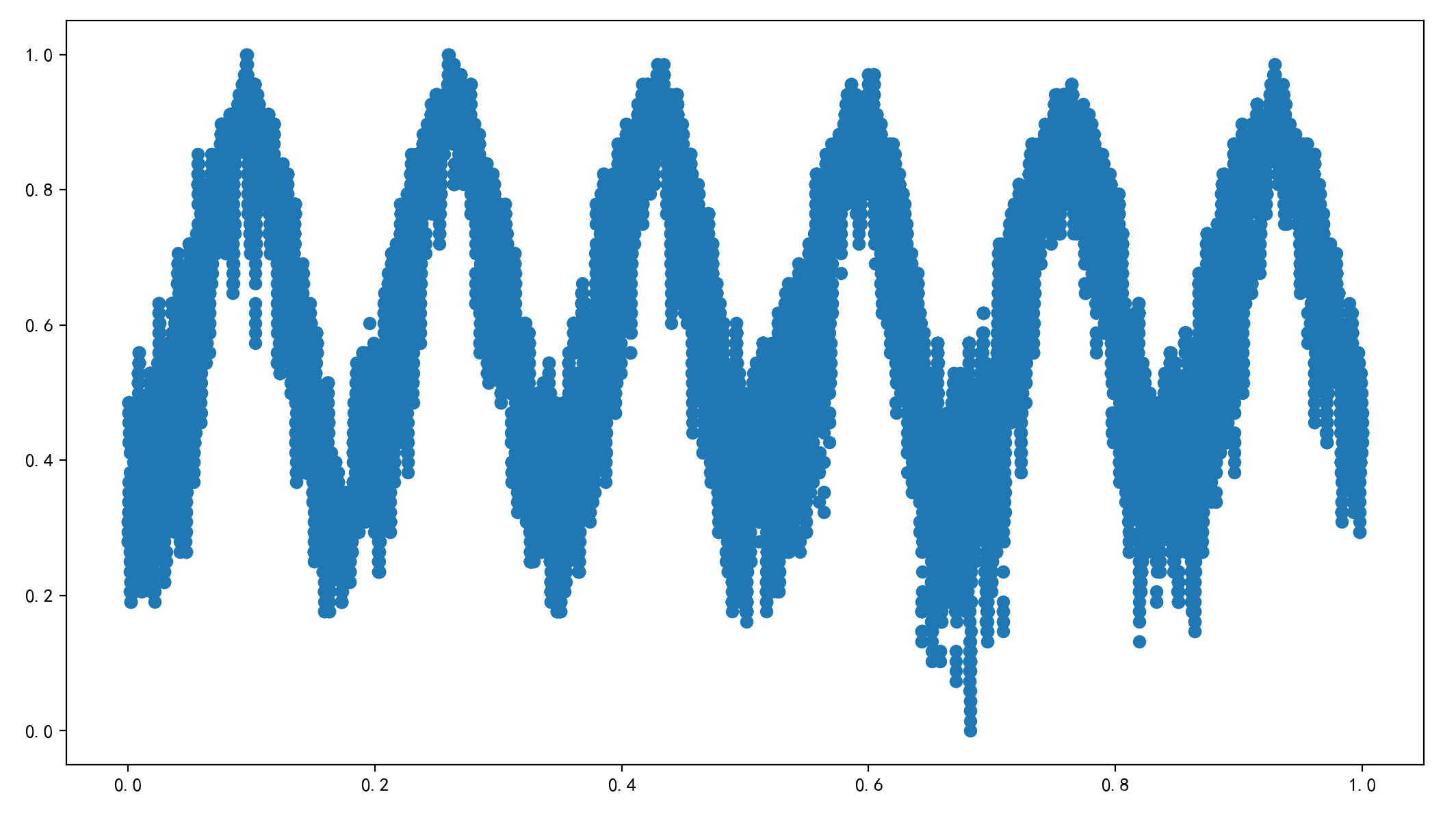


然后开始绘制0-1归一化后的两个图：

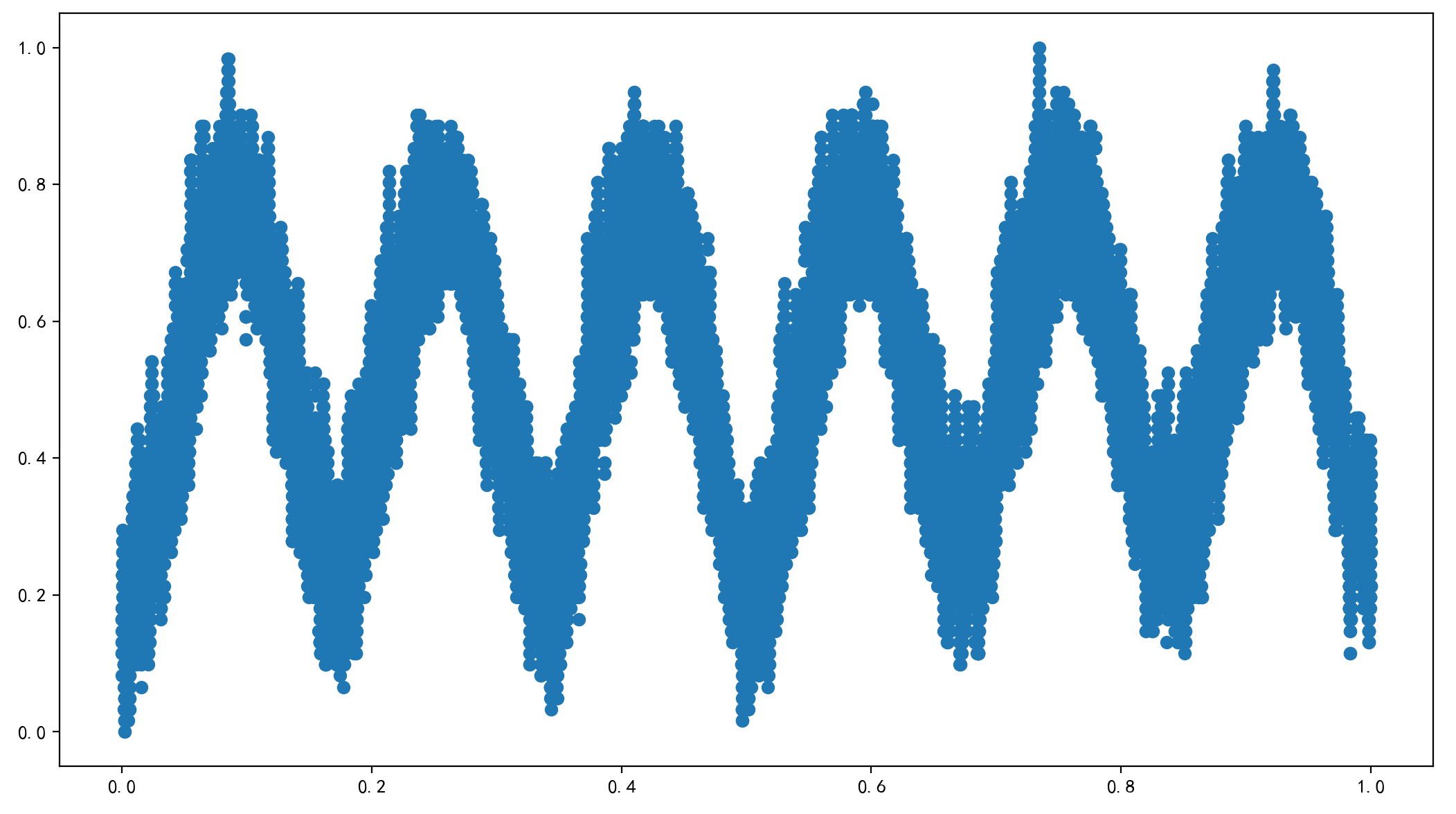


分别绘制如下：

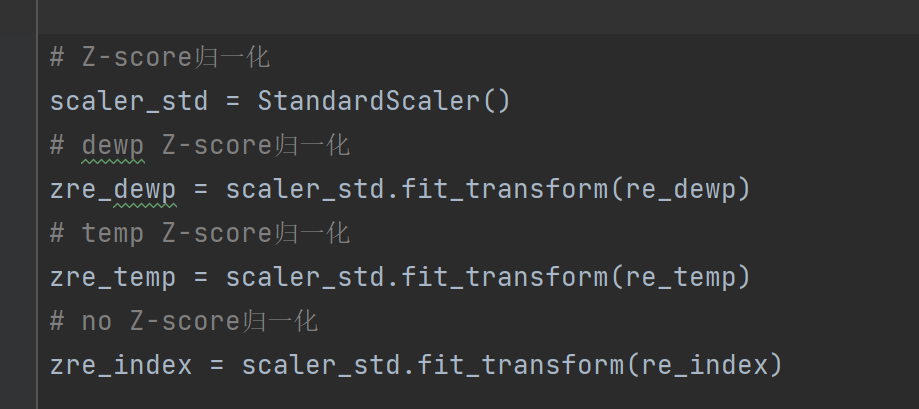
dewp的:



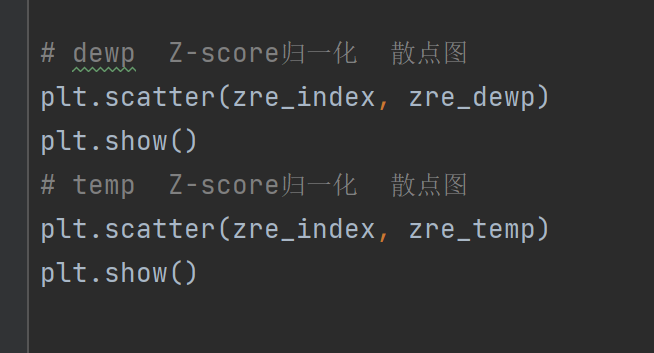
temp的：



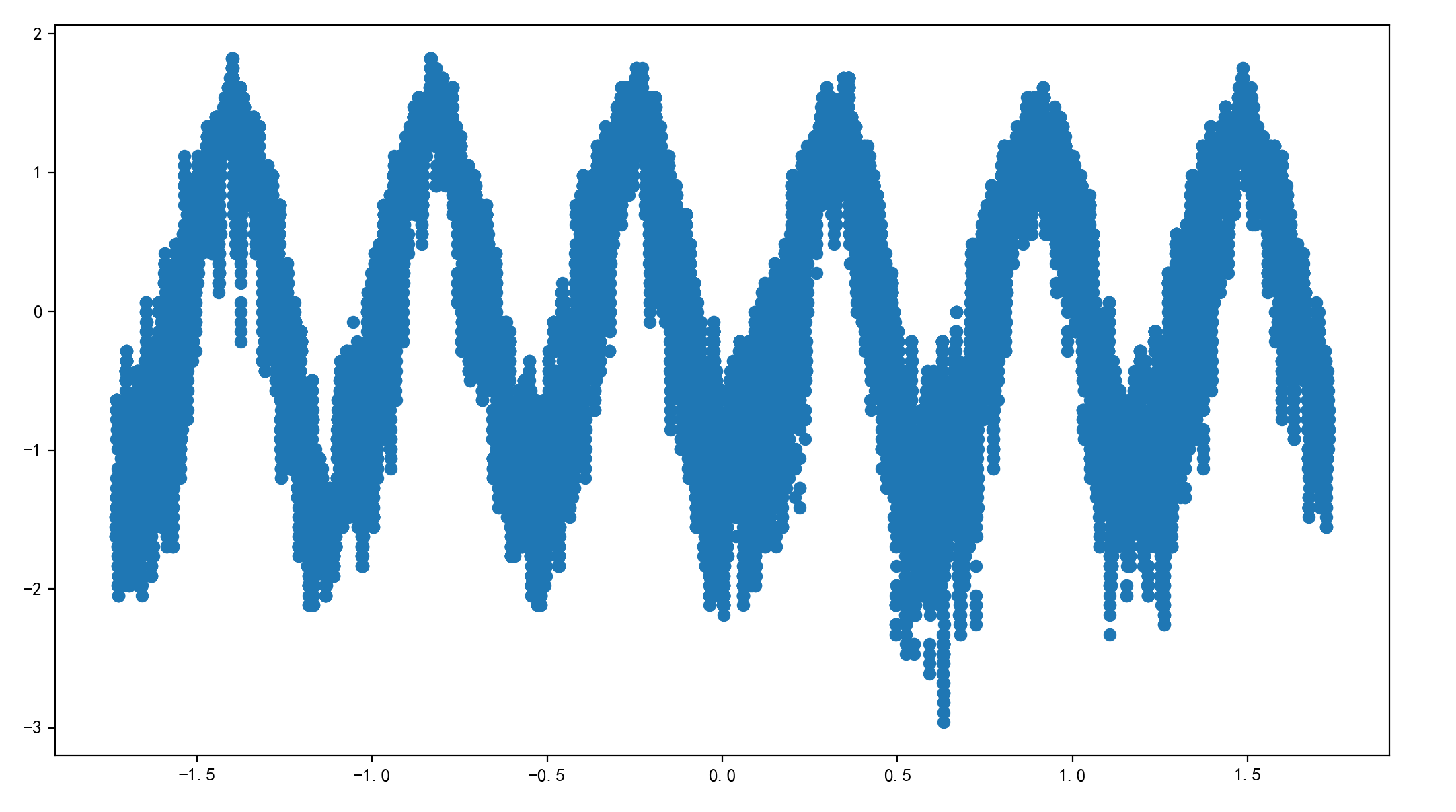
之后对数据进行Z-score归一化：



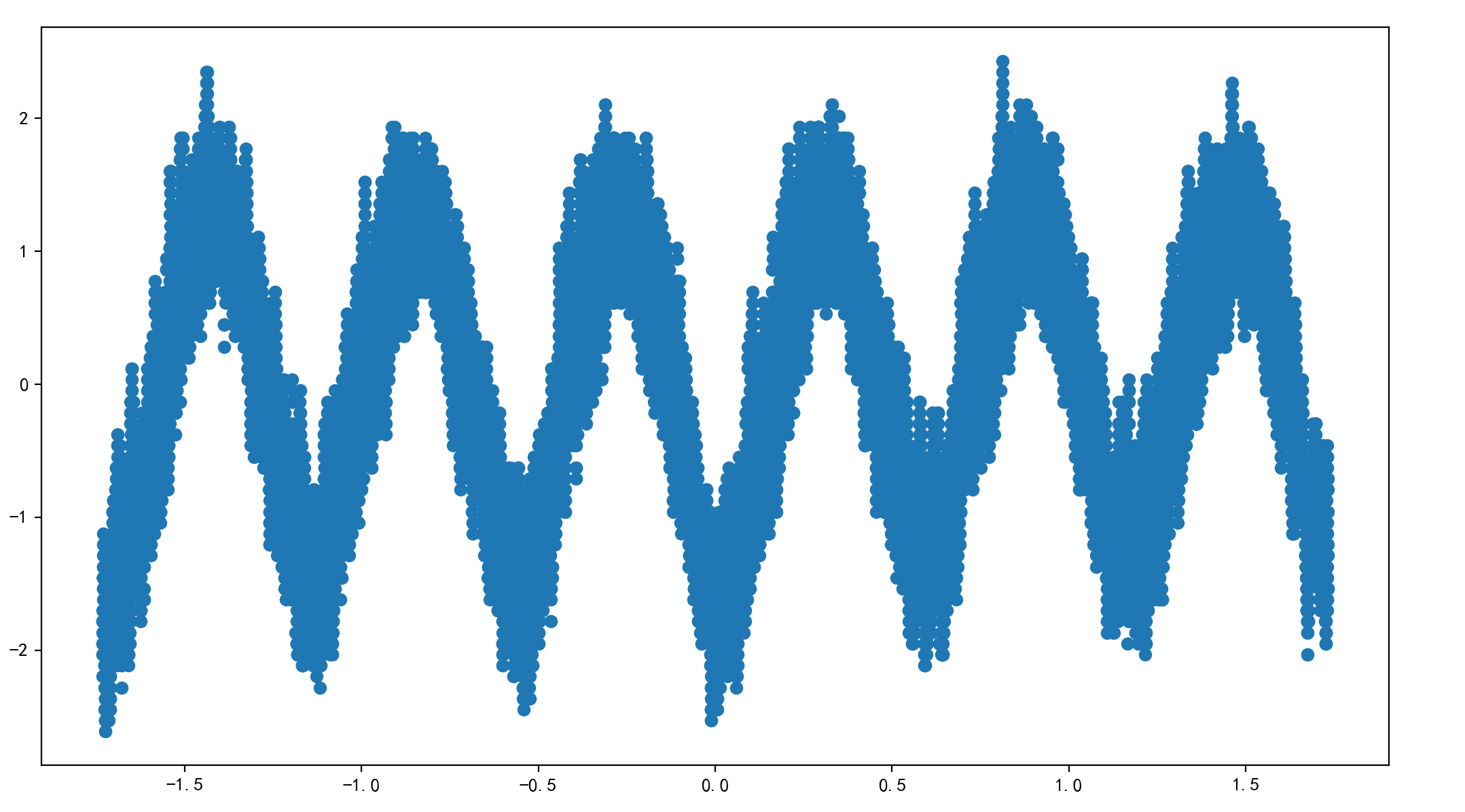
绘图：



dewp的：



temp的：



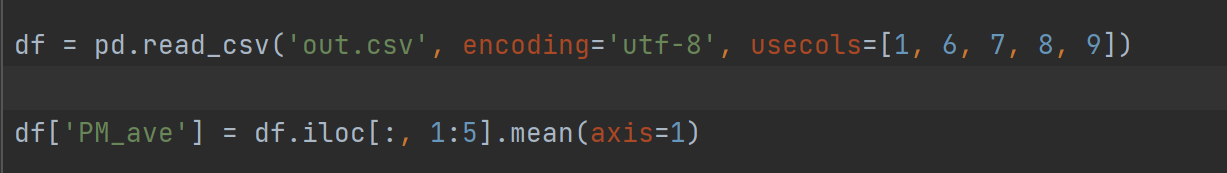
到此，第四问完成。

**5. 将北京的空气质量数据进行离散化，按照空气质量指数分级标**

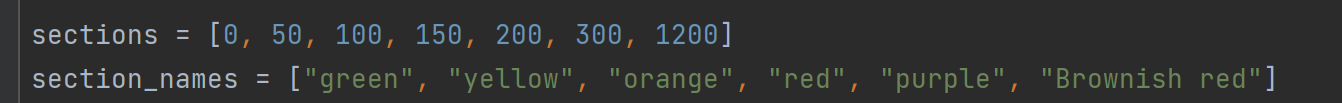
**准，计算出每个级别（或颜色值）对应的天数各有多少**

**解决第五题的代码在aqi.py里。**

**首先读入数据需要的pm的四列数据及第一列的No，并对pm数据按行求均值，保存到新的一列：**



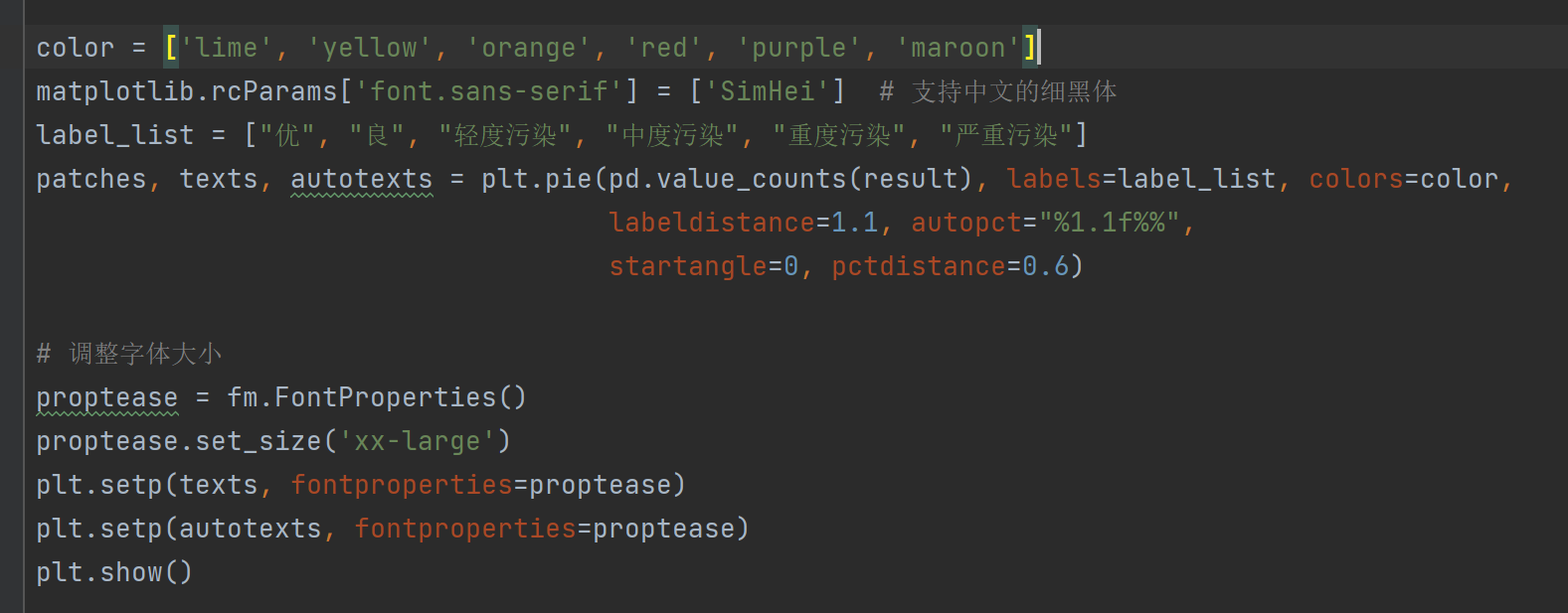
之后为离散化处理做准备：



再离散化处理数据：



然后根据处理的数据做扇形图：



得到如下图，五个问题到此均解决：

