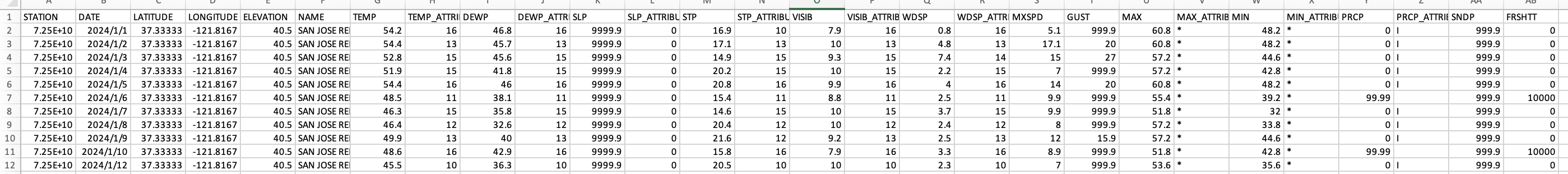
* 数据含义

NOAA全球地面每日摘要（GSOD）数据集包含气象站的每日天气摘要，每个站点一个文件，涵盖如下的一些信息

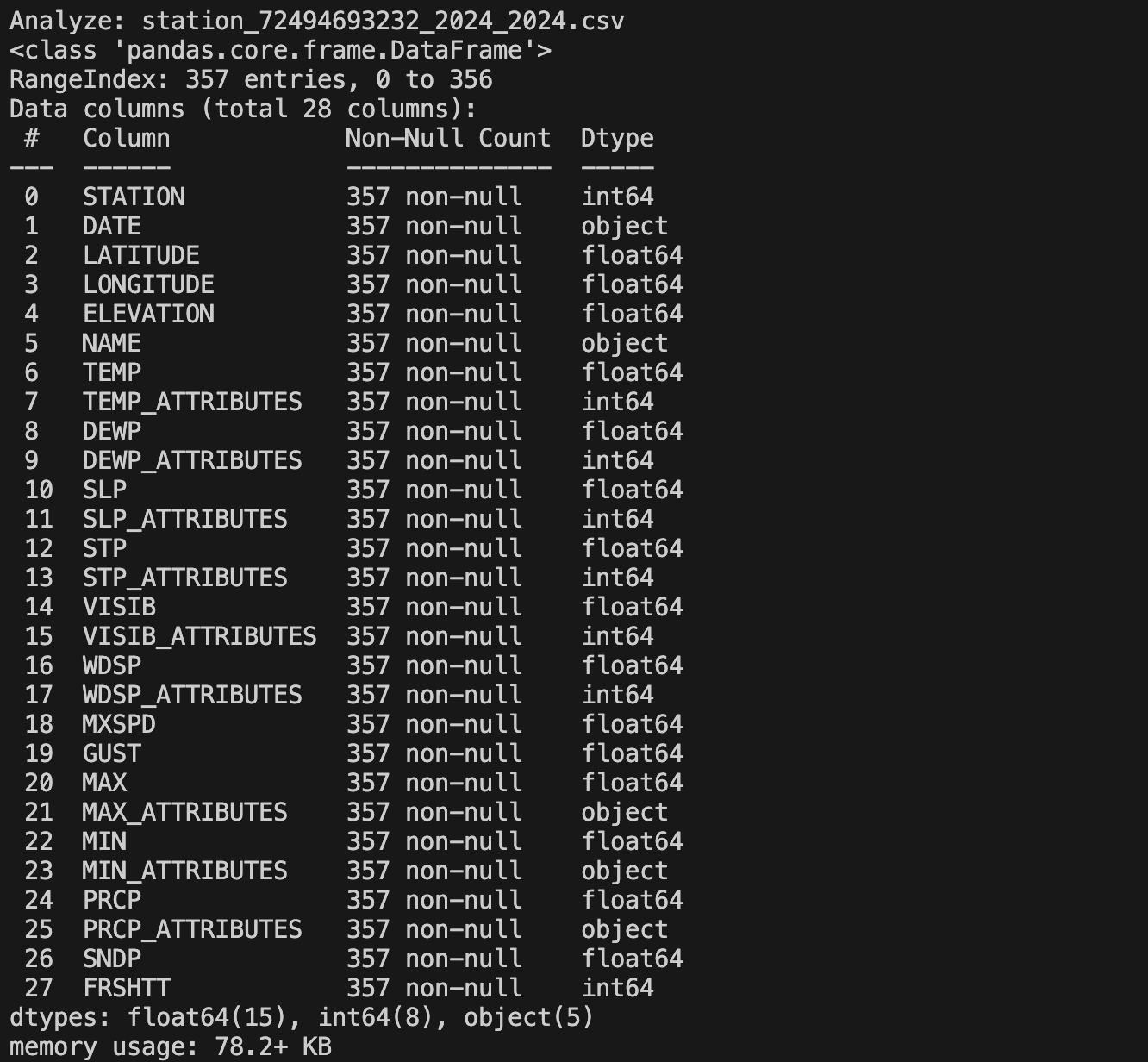
* STATION---: 站点编号（世界气象组织/DATSAV3 编号）。
* DATE: 日期，格式为YYYYMMDD。
* LATITUDE:纬度。
* LONGITUDE:经度。
* ELEVATION:海拔高度。
* NAME: 站点名称。
* TEMP: 日平均气温，单位为华氏度，精确到十分之一。缺失值表示为 9999.9。
* TEMP\_ATTRIBUTES: 用于计算平均气温的观测次数。
* DEWP: 日平均露点温度，单位为华氏度，精确到十分之一。缺失值表示为 9999.9。
* DEWP\_ATTRIBUTES: 用于计算平均露点温度的观测次数。
* SLP: 平均海平面气压，单位为毫巴，精确到十分之一。缺失值表示为 9999.9。
* SLP\_ATTRIBUTES: 用于计算平均海平面气压的观测次数。
* STP: 平均站点气压，单位为毫巴，精确到十分之一。缺失值表示为 9999.9。
* STP\_ATTRIBUTES: 用于计算平均站点气压的观测次数。
* VISIB: 平均能见度，单位为英里，精确到十分之一。缺失值表示为 999.9。
* VISIB\_ATTRIBUTES: 用于计算平均能见度的观测次数。
* WDSP: 平均风速，单位为节（海里/小时），精确到十分之一。缺失值表示为 999.9。
* WDSP\_ATTRIBUTES: 用于计算平均风速的观测次数。
* MXSPD: 最大持续风速，单位为节，精确到十分之一。缺失值表示为 999.9。
* GUST: 最大阵风风速，单位为节，精确到十分之一。缺失值表示为 999.9。
* MAX: 当日最高气温，单位为华氏度，精确到十分之一。缺失值表示为 9999.9。
* MAX\_ATTRIBUTES: 指示最高气温的获取方式（例如，来自小时数据或明确的最高气温报告）。
* MIN: 当日最低气温，单位为华氏度，精确到十分之一。缺失值表示为 9999.9。
* MIN\_ATTRIBUTES: 指示最低气温的获取方式。
* PRCP: 当日总降水量（雨和/或融化的雪），单位为英寸，精确到百分之一。“0.00”表示没有可测量的降水，“T”表示微量降水。缺失值表示为 99.99。
* PRCP\_ATTRIBUTES: 指示降水累积的时间段和其他详细信息。
* SNDP: 雪深，单位为英寸，精确到十分之一。缺失值表示为 999.9。
* FRSHTT: 指示是否发生以下天气现象：雾 (F)、雨或毛毛雨 (R)、雪或冰丸 (S)、冰雹 (H)、雷暴 (T)、龙卷风/漏斗云 (T)。“1”表示发生，“ ”（空格）表示未发生。
* 数据直观印象



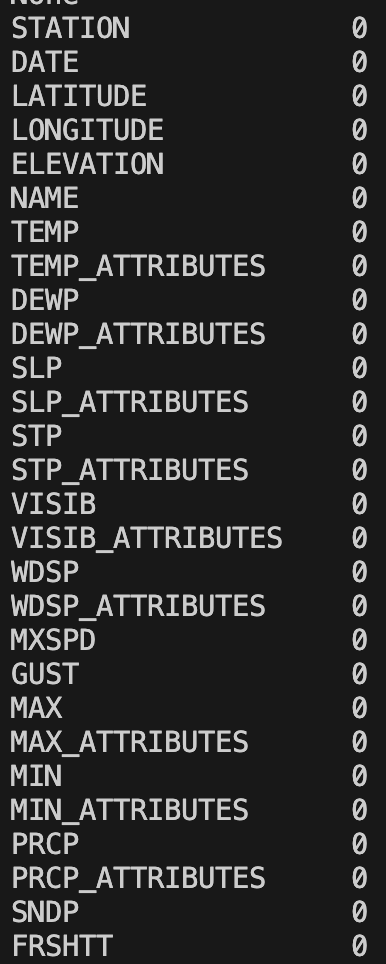
* 数据基本信息df.info()

（72494693232 san-jose站点）

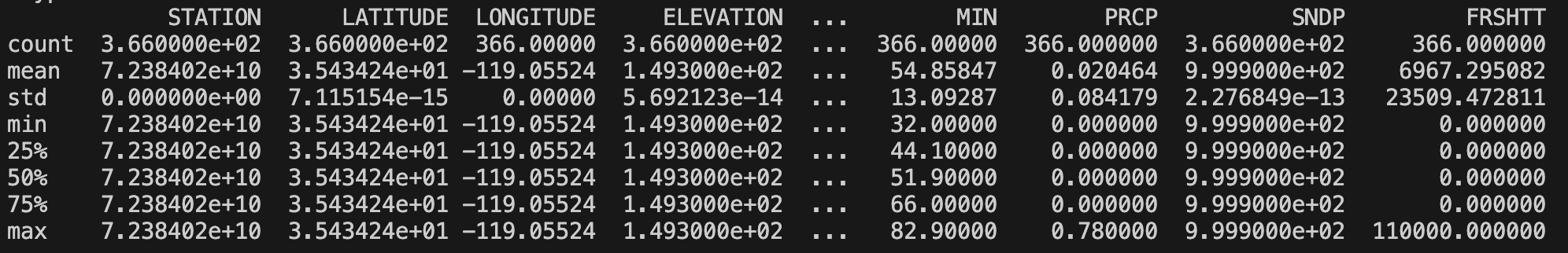
可以看到有27列数据，357行，数据完整，说明这个站点的情况整体良好。



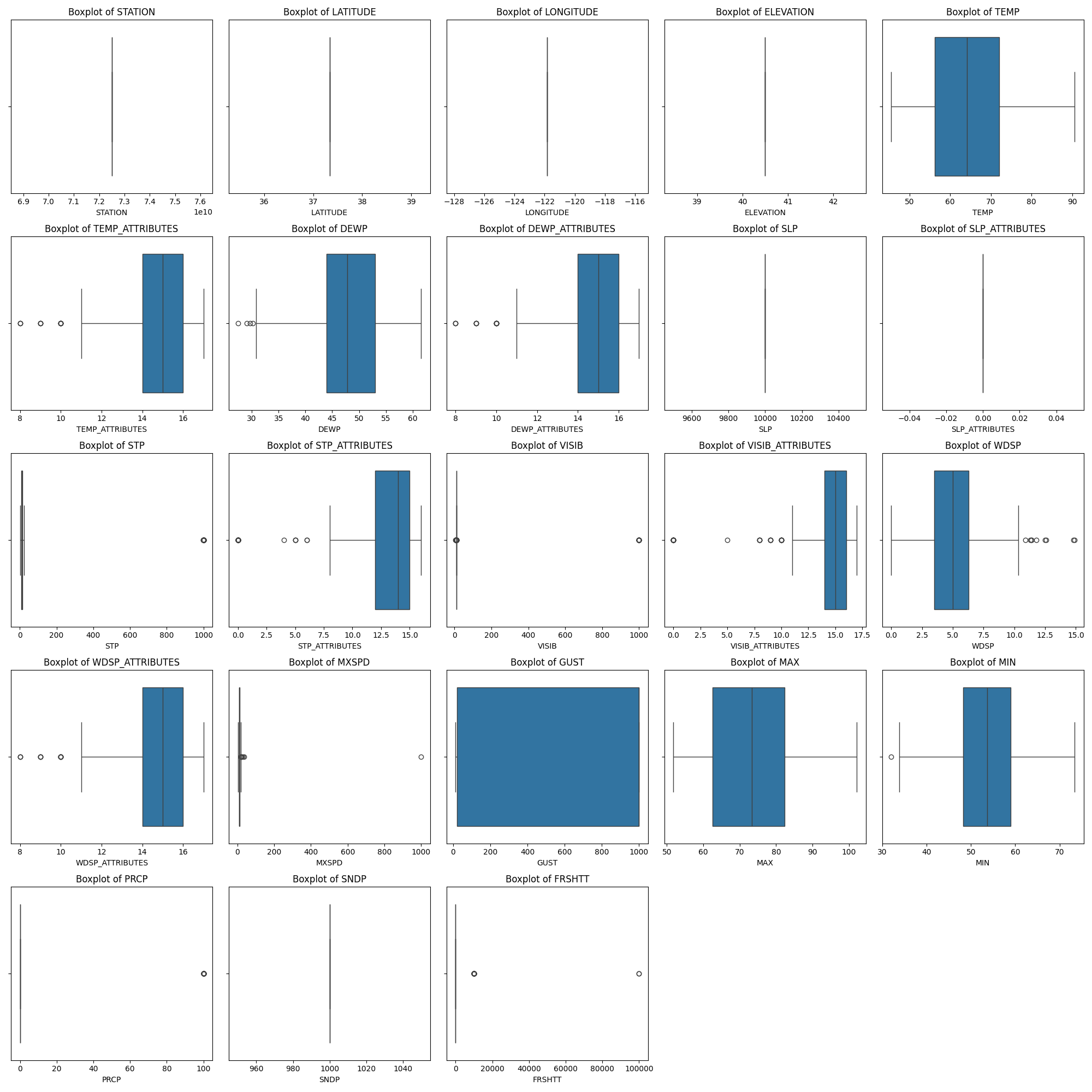
df.isnull().sum()，各列空值的统计情况，可以看到没有空值



df.describe()，整体的数据描述：计数（count）、平均值（mean）、标准差（std）、最小值（min）、四分位数（25%, 50%, 75%）和最大值（max）



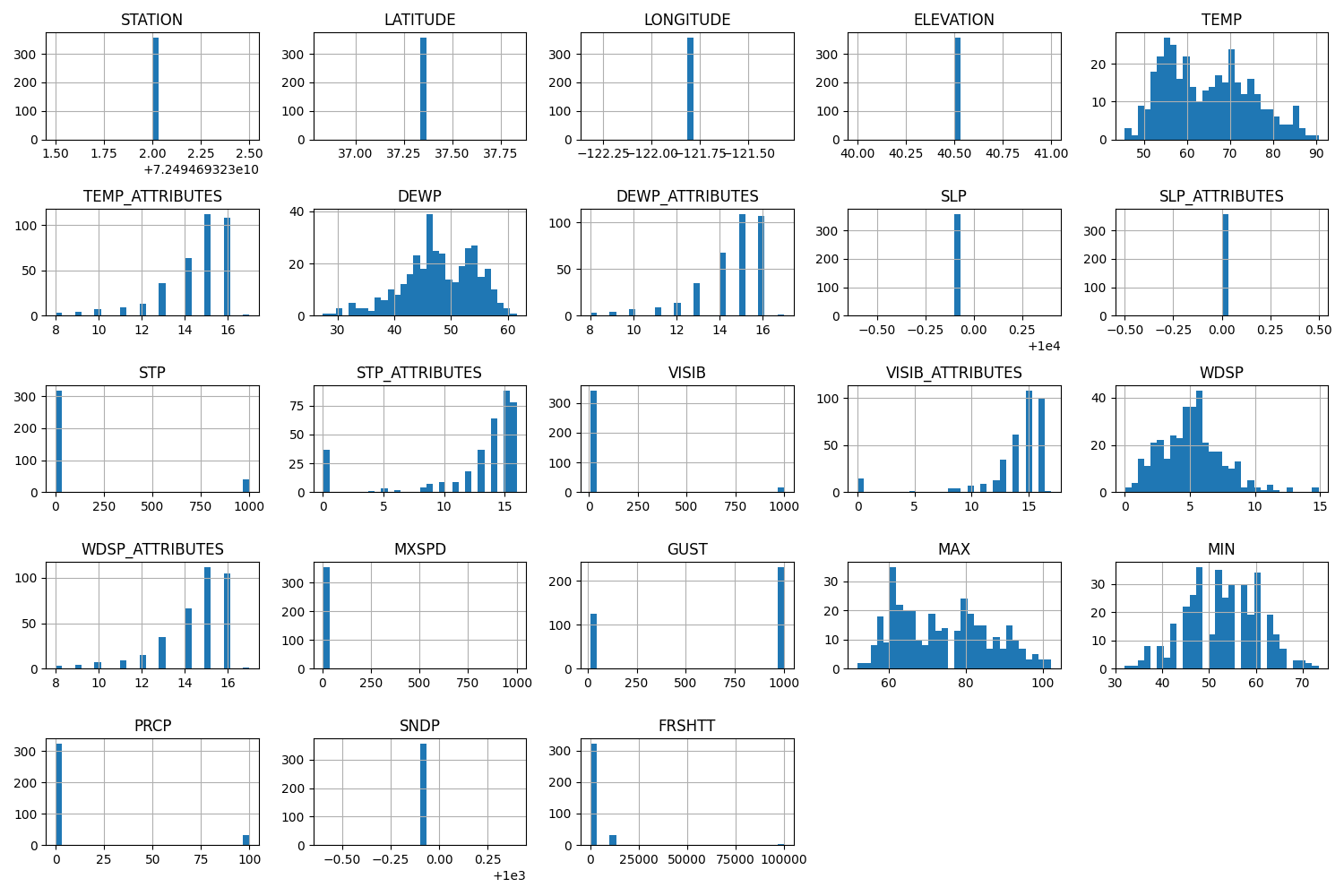
箱线图更直观看到各个数据列的数据分布：



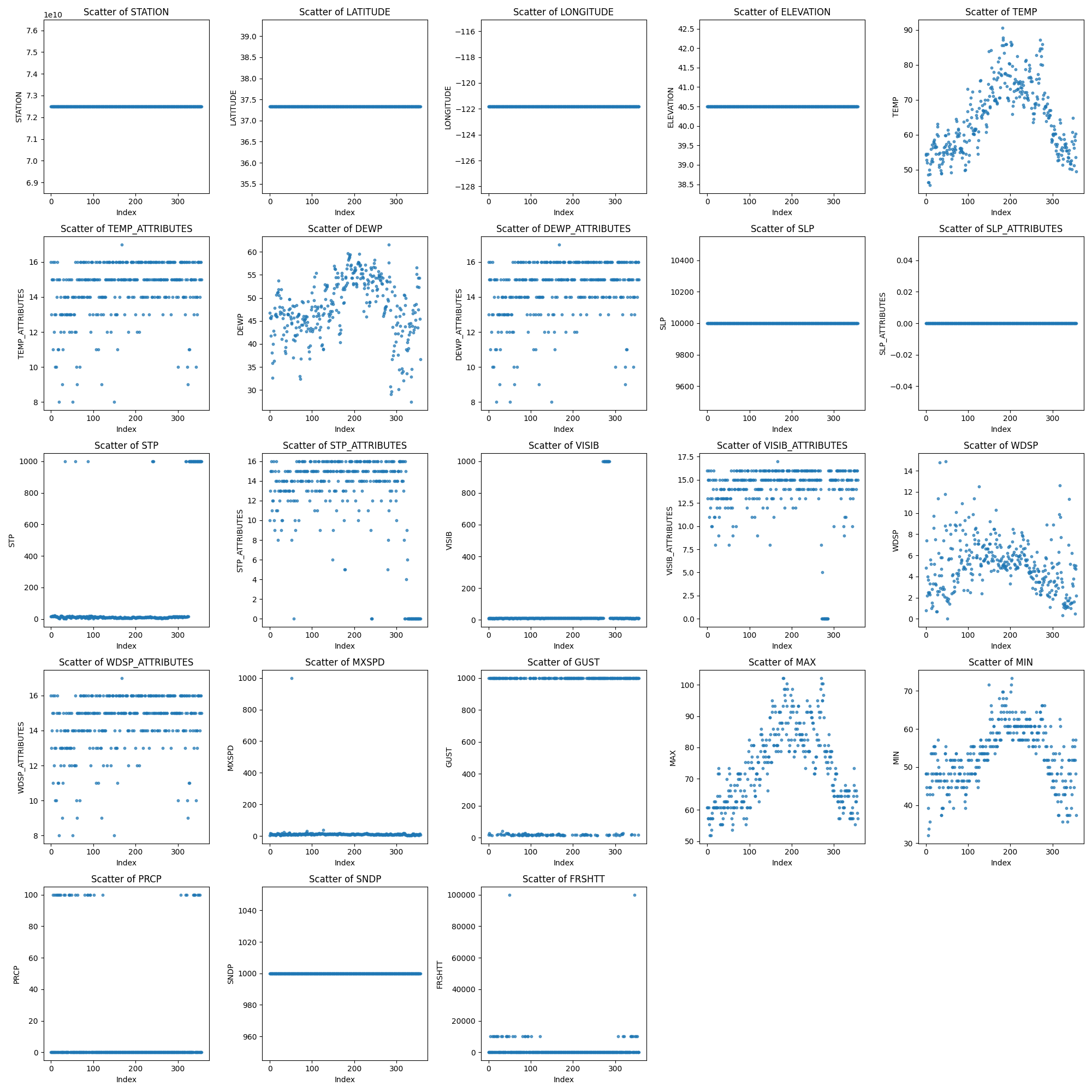
直方图

df.hist(bins=30, figsize=(15, 10))

可以看出各列数值的分布情况



散点图可以看出数据的离散情况，可以修正或者去除异常偏离的数据：



* 总结：

1. 数据概览

通过快速浏览数据集的基本信息，了解了大概数据的大概情况：共8个站点，每个站点一个文件，每个文件最多有366行数据（一年最多366天）。

1. 缺失值分析

某些站点有一段连续的时间没有数据，这些数据是不需要补充的，也不影响分析。站点中的列基本上都有数据，但是有一些数据是填充值，这些填充值跟实际值相差很大，无意义的，这行数据需要舍弃或者通过插值的方法补充上去。

1. 数据分布分析

我们利用直方图、箱线图等可视化工具分析数值型变量的分布情况，发现有不少数值属于偏态分布，可能需要进行转换来减少偏差；有一些列根据特征工程需要舍弃的，就不做过多分析。

1. 异常值检测

使用散点图检测异常值，可以识别出的异常值，根据站点的散点图，数据基本上比较正常。

EDA分析结果

1. 根据EPA对天气与空气质量的关系说明（光照、高温、降雨、风速等），GSOD的数据列含义和数据的基本情况进行分析

|  |  |  |  |
| --- | --- | --- | --- |
| 列 | 含义 | 特征相关性 | 处理方法 |
| STATION | 站点编号（世界气象组织/DATSAV3 编号）。 | 对AQI无意义 | 留存，训练时不用 |
| DATE | 日期，格式为YYYYMMDD。 | 对AQI无意义 | 留存，训练时不用 |
| LATITUDE | 纬度 | 不变值对AQI无意义 | 删除 |
| LONGITUDE | 经度 | 不变值对AQI无意义 | 删除 |
| ELEVATION | 海拔 | 不变值对AQI无意义 | 删除 |
| NAME | 站点名 | 对AQI无意义 | 留存，训练时不用 |
| TEMP | 日平均气温，单位为华氏度，精确到十分之一。缺失值表示为 9999.9。 | 重要 | 保留，缺失值需要替换，可以采用前后一日的平均值，前后缺失则找更远的日期 |
| TEMP\_ATTRIBUTES | 用于计算平均气温的观测次数。 | 对AQI无意义 | 删除 |
| DEWP | 日平均露点温度，单位为华氏度，精确到十分之一。缺失值表示为 9999.9。 | 重要 | 保留，缺失值需要替换，可以采用前后一日的平均值，前后缺失则找更远的日期 |
| DEWP\_ATTRIBUTES | 用于计算平均露点温度的观测次数。 | 对AQI无意义 | 删除 |
| SLP | 平均海平面气压，单位为毫巴，精确到十分之一。缺失值表示为 9999.9。 | 对AQI无意义 | 删除 |
| SLP\_ATTRIBUTES | 用于计算平均海平面气压的观测次数。 | 对AQI无意义 | 删除 |
| STP | 平均站点气压，单位为毫巴，精确到十分之一。缺失值表示为 9999.9。 | 有意义 | 保留，缺失值需要替换，可以采用前后一日的平均值，前后缺失则找更远的日期 |
| STP\_ATTRIBUTES | 用于计算平均站点气压的观测次数。 | 对AQI无意义 | 删除 |
| VISIB | 平均能见度，单位为英里，精确到十分之一。缺失值表示为 999.9。 | 有意义 | 保留，缺失值需要替换，可以采用前后一日的平均值，前后缺失则找更远的日期 |
| VISIB\_ATTRIBUTES | 用于计算平均能见度的观测次数。 | 对AQI无意义 | 删除 |
| WDSP | 平均风速，单位为节（海里/小时），精确到十分之一。缺失值表示为 999.9。 | 有意义 | 保留，缺失值需要替换，可以采用前后一日的平均值，前后缺失则找更远的日期 |
| WDSP\_ATTRIBUTES | 用于计算平均风速的观测次数。 | 对AQI无意义 | 删除 |
| MXSPD | 最大持续风速，单位为节，精确到十分之一。缺失值表示为 999.9。 | 有意义 | 保留，缺失值需要替换，可以采用前后一日的平均值，前后缺失则找更远的日期 |
| GUST | 最大阵风风速，单位为节，精确到十分之一。缺失值表示为 999.9。 | 有意义 | 缺失值太多，删除 |
| MAX: | 当日最高气温，单位为华氏度，精确到十分之一。缺失值表示为 9999.9。 | 有意义 | 保留，缺失值需要替换，可以采用前后一日的平均值，前后缺失则找更远的日期 |
| MAX\_ATTRIBUTES | 指示最高气温的获取方式（例如，来自小时数据或明确的最高气温报告）。 | 对AQI无意义 | 删除 |
| MIN | 当日最低气温，单位为华氏度，精确到十分之一。缺失值表示为 9999.9。 | 有意义 | 保留，缺失值需要替换，可以采用前后一日的平均值，前后缺失则找更远的日期 |
| MIN\_ATTRIBUTES | 指示最低气温的获取方式。 | 对AQI无意义 | 删除 |
| PRCP | 当日总降水量（雨和/或融化的雪），单位为英寸，精确到百分之一。“0.00”表示没有可测量的降水，“T”表示微量降水。缺失值表示为 99.99。 | 有意义 | 保留，缺失值使用0来替代，因为超过一半的情况是0 |
| PRCP\_ATTRIBUTES | 指示降水累积的时间段和其他详细信息。 | 对AQI无意义 | 删除 |
| SNDP: | 雪深，单位为英寸，精确到十分之一。缺失值表示为 999.9。 | 无意义 | 删除（值几乎都缺失） |
| FRSHTT | 指示是否发生以下天气现象：雾 (F)、雨或毛毛雨 (R)、雪或冰丸 (S)、冰雹 (H)、雷暴 (T)、龙卷风/漏斗云 (T)。“1”表示发生，“ ”（空格）表示未发生。 | 有意义 | 数据异常，删除 |

1. 保留列图：

|  |  |  |  |
| --- | --- | --- | --- |
| 列 | 含义 | 特征相关性 | 处理方法 |
| STATION | 站点编号（世界气象组织/DATSAV3 编号）。 | 对AQI无意义 | 留存，训练时不用 |
| DATE | 日期，格式为YYYYMMDD。 | 对AQI无意义 | 留存，训练时不用 |
| TEMP | 日平均气温，单位为华氏度，精确到十分之一。缺失值表示为 9999.9。 | 重要 | 保留，缺失值需要替换，可以采用前后一日的平均值，前后缺失则找更远的日期 |
| NAME | 站点名 | 对AQI无意义 | 留存，训练时不用，缺失值使用前后的值进行填充即可 |
| DEWP | 日平均露点温度，单位为华氏度，精确到十分之一。缺失值表示为 9999.9。 | 重要 | 保留，缺失值需要替换，可以采用前后一日的平均值，前后缺失则找更远的日期 |
| STP | 平均站点气压，单位为毫巴，精确到十分之一。缺失值表示为 9999.9。 | 有意义 | 保留，缺失值需要替换，可以采用前后一日的平均值，前后缺失则找更远的日期 |
| VISIB | 平均能见度，单位为英里，精确到十分之一。缺失值表示为 999.9。 | 有意义 | 保留，缺失值需要替换，可以采用前后一日的平均值，前后缺失则找更远的日期 |
| WDSP | 平均风速，单位为节（海里/小时），精确到十分之一。缺失值表示为 999.9。 | 有意义 | 保留，缺失值需要替换，可以采用前后一日的平均值，前后缺失则找更远的日期 |
| MXSPD | 最大持续风速，单位为节，精确到十分之一。缺失值表示为 999.9。 | 有意义 | 保留，缺失值需要替换，可以采用前后一日的平均值，前后缺失则找更远的日期 |
| MAX: | 当日最高气温，单位为华氏度，精确到十分之一。缺失值表示为 9999.9。 | 有意义 | 保留，缺失值需要替换，可以采用前后一日的平均值，前后缺失则找更远的日期 |
| MIN | 当日最低气温，单位为华氏度，精确到十分之一。缺失值表示为 9999.9。 | 有意义 | 保留，缺失值需要替换，可以采用前后一日的平均值，前后缺失则找更远的日期 |
| PRCP | 当日总降水量（雨和/或融化的雪），单位为英寸，精确到百分之一。“0.00”表示没有可测量的降水，“T”表示微量降水。缺失值表示为 99.99。 | 有意义 | 保留，缺失值使用0来替代，因为超过一半的情况是0 |

AQI的研究来看，季节数据对对AQI来说可能有实际意义，所以我们在数据中会根据日期情况，加入一列月份数据。

* 数据的实际处理（聚合、清洗、筛选等）：

1. 去掉不需要的列LATITUDE，ELEVATION，LONGITUDE，TEMP\_ATTRIBUTES，DEWP\_ATTRIBUTES，SLP，SLP\_ATTRIBUTES，STP\_ATTRIBUTES，VISIB\_ATTRIBUTES，WDSP\_ATTRIBUTES，GUST，MAX\_ATTRIBUTES，MIN\_ATTRIBUTES，PRCP，PRCP\_ATTRIBUTES，SNDP，FRSHTT
2. 处理TEMP列的日平均气温数据，如果数据大于200，则使用最近前一条和最近后一条的平均值填充，如果没有前一条或者后一条，则只用有的一条填充，如果前后都没有，则删除这个日期的这一行数据。用来填充的前一条或者后一条数据都不能大于200。
3. 处理DEWP列的日平均露点温数据，如果数据大于200，则使用最近前一条和最近后一条的平均值填充，如果没有前一条或者后一条，则只用有的一条填充，如果前后都没有，则删除这个日期的这一行数据。用来填充的前一条或者后一条数据都不能大于200。打印填充或者删除日志信息
4. 处理STP列的平均站点气压数据，如果数据为9999.9，则使用最近前一条和最近后一条的平均值填充，如果没有前一条或者后一条，则只用有的一条填充，如果前后都没有，则删除这个日期的这一行数据。用来填充的前一条或者后一条数据都不能是9999.9。
5. 处理VISIB列的平均能见度数据，如果数据为999.9，则使用最近前一条和最近后一条的平均值填充，如果没有前一条或者后一条，则只用有的一条填充，如果前后都没有，则删除这个日期的这一行数据。用来填充的前一条或者后一条数据都不能是999.9。
6. 处理WDSP列的平均风速数据，如果数据为999.9，则使用最近前一条和最近后一条的平均值填充，如果没有前一条或者后一条，则只用有的一条填充，如果前后都没有，则删除这个日期的这一行数据。用来填充的前一条或者后一条数据都不能是999.9。
7. 处理MXSPD列的最大持续风速数据，如果数据为999.9，则使用最近前一条和最近后一条的平均值填充，如果没有前一条或者后一条，则只用有的一条填充，如果前后都没有，则删除这个日期的这一行数据。用来填充的前一条或者后一条数据都不能是999.9。
8. 处理MAX列的当日最高气温数据，如果数据为9999.9，则使用最近前一条和最近后一条的平均值填充，如果没有前一条或者后一条，则只用有的一条填充，如果前后都没有，则删除这个日期的这一行数据。用来填充的前一条或者后一条数据都不能是9999.9。
9. 处理MIN列的当日最低气温数据，如果数据为9999.9，则使用最近前一条和最近后一条的平均值填充，如果没有前一条或者后一条，则只用有的一条填充，如果前后都没有，则删除这个日期的这一行数据。用来填充的前一条或者后一条数据都不能是9999.9。
10. 处理PRCP列的当日总降水量数据，如果数据为99.9，设置为0.00。
11. 处理后的noaa数据文件上加上一列month，数据为1,2,3,4直到12，根据date数据列来生成。
12. 处理后检查数据文件，如果一行只有AQI数据或者一行没有AQI数据，则删除此行。