rainfall-prediction

February 3, 2024

[1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

from sklearn import preprocessing
import scipy.stats as stats
from sklearn.model_selection import train_test_split
from collections import Counter
from sklearn.metrics import accuracy_score, confusion_matrix,␣
 ↪classification_report
from sklearn import metrics
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier
from sklearn.svm import SVC
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import GaussianNB
from sklearn.neighbors import KNeighborsClassifier
```

[2]:
```
df = pd.read_csv('weatherAUS.csv')
pd.set_option('display.max_columns',None)
```

[3]:
```
df
```

[3]:
```
              Date Location  MinTemp  MaxTemp  Rainfall  Evaporation  \
0       2008-12-01   Albury     13.4     22.9       0.6          NaN
1       2008-12-02   Albury      7.4     25.1       0.0          NaN
2       2008-12-03   Albury     12.9     25.7       0.0          NaN
3       2008-12-04   Albury      9.2     28.0       0.0          NaN
4       2008-12-05   Albury     17.5     32.3       1.0          NaN
…                …        …      …        …        …
142188  2017-06-20    Uluru      3.5     21.8       0.0          NaN
142189  2017-06-21    Uluru      2.8     23.4       0.0          NaN
142190  2017-06-22    Uluru      3.6     25.3       0.0          NaN
142191  2017-06-23    Uluru      5.4     26.9       0.0          NaN
142192  2017-06-24    Uluru      7.8     27.0       0.0          NaN
```

```
        Sunshine WindGustDir  WindGustSpeed WindDir9am WindDir3pm  \
0            NaN           W           44.0          W        WNW
1            NaN         WNW           44.0        NNW        WSW
2            NaN         WSW           46.0          W        WSW
3            NaN          NE           24.0         SE          E
4            NaN           W           41.0        ENE         NW
…          …         …            …        …        …
142188       NaN           E           31.0        ESE          E
142189       NaN           E           31.0         SE        ENE
142190       NaN         NNW           22.0         SE          N
142191       NaN           N           37.0         SE        WNW
142192       NaN          SE           28.0        SSE          N

        WindSpeed9am  WindSpeed3pm  Humidity9am  Humidity3pm  Pressure9am  \
0               20.0          24.0         71.0         22.0       1007.7
1                4.0          22.0         44.0         25.0       1010.6
2               19.0          26.0         38.0         30.0       1007.6
3               11.0           9.0         45.0         16.0       1017.6
4                7.0          20.0         82.0         33.0       1010.8
…              …           …          …          …          …
142188          15.0          13.0         59.0         27.0       1024.7
142189          13.0          11.0         51.0         24.0       1024.6
142190          13.0           9.0         56.0         21.0       1023.5
142191           9.0           9.0         53.0         24.0       1021.0
142192          13.0           7.0         51.0         24.0       1019.4

        Pressure3pm  Cloud9am  Cloud3pm  Temp9am  Temp3pm RainToday  RISK_MM  \
0            1007.1       8.0       NaN     16.9     21.8        No      0.0
1            1007.8       NaN       NaN     17.2     24.3        No      0.0
2            1008.7       NaN       2.0     21.0     23.2        No      0.0
3            1012.8       NaN       NaN     18.1     26.5        No      1.0
4            1006.0       7.0       8.0     17.8     29.7        No      0.2
…             …       …       …      …      …       …      …
142188       1021.2       NaN       NaN      9.4     20.9        No      0.0
142189       1020.3       NaN       NaN     10.1     22.4        No      0.0
142190       1019.1       NaN       NaN     10.9     24.5        No      0.0
142191       1016.8       NaN       NaN     12.5     26.1        No      0.0
142192       1016.5       3.0       2.0     15.1     26.0        No      0.0

       RainTomorrow
0                No
1                No
2                No
3                No
4                No
…             …
142188           No
```

```
142189           No
142190           No
142191           No
142192           No

[142193 rows x 24 columns]
```

[4]:

```
numerical_feature = [feature for feature in df.columns if df[feature].dtypes != 'O']
discrete_feature = [feature for feature in numerical_feature if len(df[feature].unique())<25]
continuous_feature = [feature for feature in numerical_feature if feature not in discrete_feature]
categorical_feature = [feature for feature in df.columns if feature not in numerical_feature]

print('Numerical Feature Count {}'.format(len(numerical_feature)))
print('Discrete Feature Count {}'.format(len(discrete_feature)))
print('Continuous Feature Count {}'.format(len(continuous_feature)))
print('Categorical Feature Count {}'.format(len(categorical_feature)))
```

```
Numerical Feature Count 17
Discrete Feature Count 2
Continuous Feature Count 15
Categorical Feature Count 7
```

[5]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 142193 entries, 0 to 142192
Data columns (total 24 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   Date           142193 non-null  object
 1   Location       142193 non-null  object
 2   MinTemp        141556 non-null  float64
 3   MaxTemp        141871 non-null  float64
 4   Rainfall       140787 non-null  float64
 5   Evaporation    81350 non-null   float64
 6   Sunshine       74377 non-null   float64
 7   WindGustDir    132863 non-null  object
 8   WindGustSpeed  132923 non-null  float64
 9   WindDir9am     132180 non-null  object
 10  WindDir3pm     138415 non-null  object
 11  WindSpeed9am   140845 non-null  float64
 12  WindSpeed3pm   139563 non-null  float64
 13  Humidity9am    140419 non-null  float64
 14  Humidity3pm    138583 non-null  float64
 15  Pressure9am    128179 non-null  float64
 16  Pressure3pm    128212 non-null  float64
 17  Cloud9am       88536 non-null   float64
 18  Cloud3pm       85099 non-null   float64
 19  Temp9am        141289 non-null  float64
 20  Temp3pm        139467 non-null  float64
 21  RainToday      140787 non-null  object
 22  RISK_MM        142193 non-null  float64
 23  RainTomorrow   142193 non-null  object
dtypes: float64(17), object(7)
memory usage: 26.0+ MB
```

[7]:

```
null_value = df.isnull().sum()*100/len(df)
null_value
```

[7]:

```
Date              0.000000
Location          0.000000
MinTemp           0.447983
MaxTemp           0.226453
Rainfall          0.988797
Evaporation      42.789026
Sunshine         47.692924
WindGustDir       6.561504
WindGustSpeed     6.519308
WindDir9am        7.041838
WindDir3pm        2.656952
WindSpeed9am      0.948007
WindSpeed3pm      1.849599
Humidity9am       1.247600
Humidity3pm       2.538803
Pressure9am       9.855619
Pressure3pm       9.832411
Cloud9am         37.735332
Cloud3pm         40.152469
Temp9am           0.635756
Temp3pm           1.917113
RainToday         0.988797
RISK_MM           0.000000
RainTomorrow      0.000000
dtype: float64
```

[8]:

```
print(numerical_feature)
```

```
['MinTemp', 'MaxTemp', 'Rainfall', 'Evaporation', 'Sunshine', 'WindGustSpeed',
'WindSpeed9am', 'WindSpeed3pm', 'Humidity9am', 'Humidity3pm', 'Pressure9am',
'Pressure3pm', 'Cloud9am', 'Cloud3pm', 'Temp9am', 'Temp3pm', 'RISK_MM']
```

[9]:

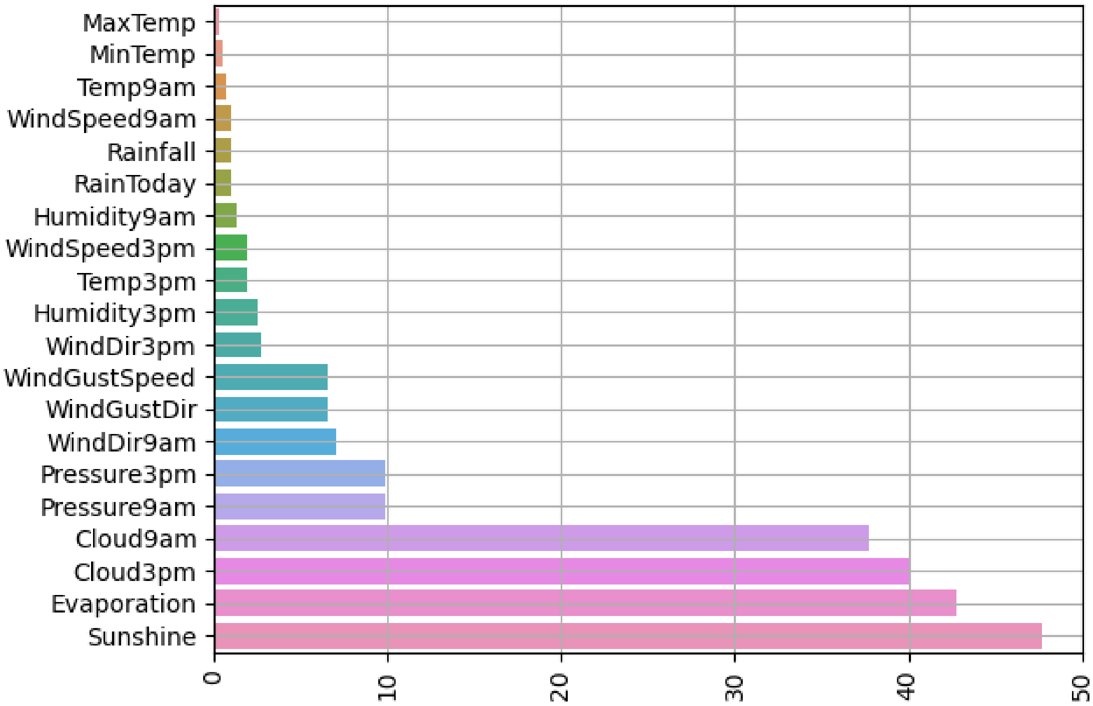
```
def missing_value(df):
    value_percent = 100*df.isnull().sum()/len(df)
    value_percent = value_percent[value_percent > 0].sort_values()
    return value_percent
```

[10]:

```
value_percent = missing_value(df)
```

[11]:

```
sns.barplot(y = value_percent.index, x = value_percent)
plt.xticks(rotation=90)
plt.grid()
plt.show()
```



[25]:

```
df['Sunshine'] = df['Sunshine'].fillna(df['Sunshine'].mean())
df['Evaporation'] = df['Evaporation'].fillna(df['Evaporation'].mean())
df['Cloud3pm'] = df['Cloud3pm'].fillna(df['Cloud3pm'].mean())
df['Cloud9am'] = df['Cloud9am'].fillna(df['Cloud9am'].mean())
```

[30]:

```
df = df.dropna()
df.isnull().sum()
```

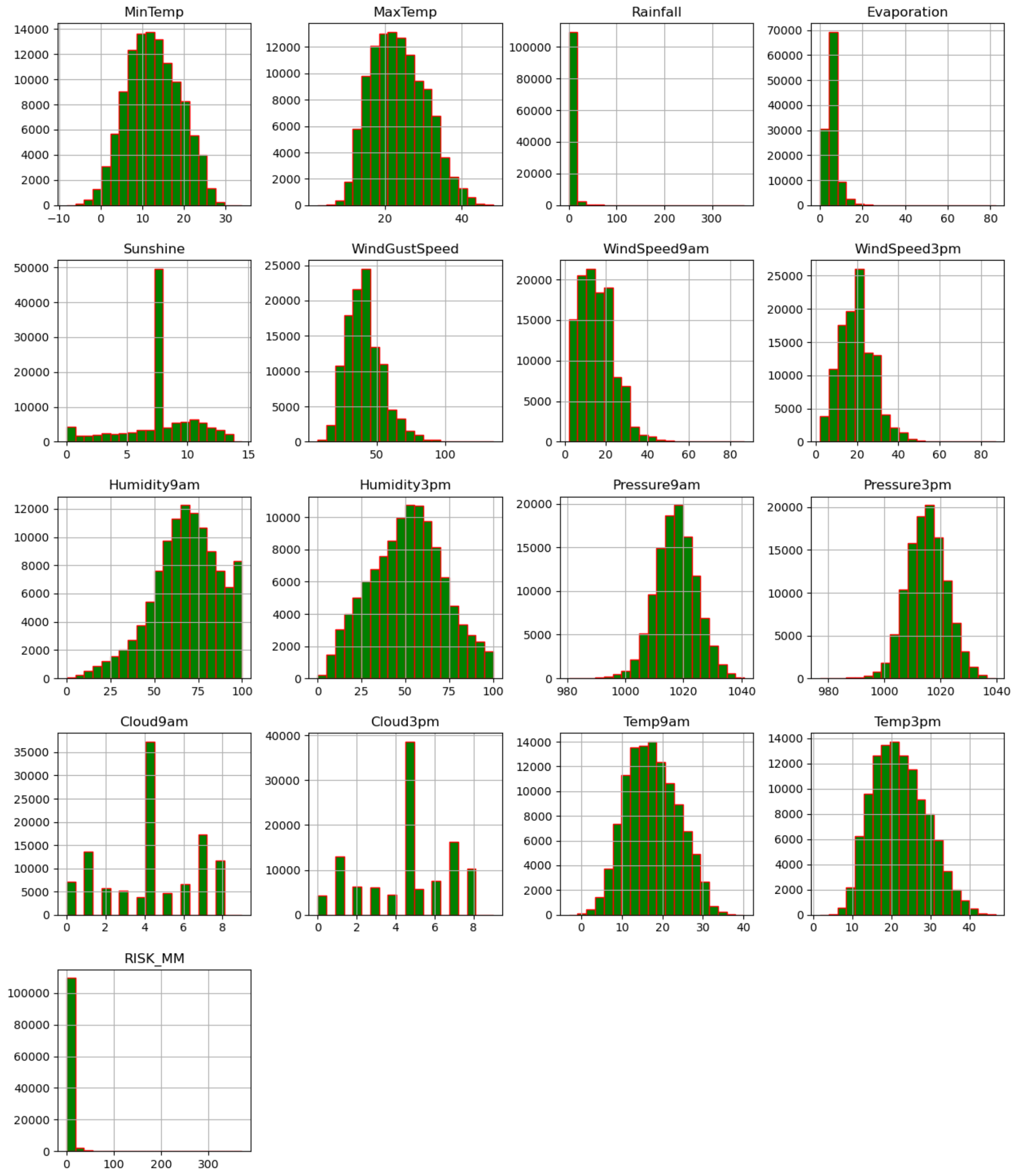
[30]:

```
Date             0
Location         0
MinTemp          0
MaxTemp          0
```

```
Rainfall         0
Evaporation      0
Sunshine         0
WindGustDir      0
WindGustSpeed    0
WindDir9am       0
WindDir3pm       0
WindSpeed9am     0
WindSpeed3pm     0
Humidity9am      0
Humidity3pm      0
Pressure9am      0
Pressure3pm      0
Cloud9am         0
Cloud3pm         0
Temp9am          0
Temp3pm          0
RainToday        0
RISK_MM          0
RainTomorrow     0
dtype: int64
```

[38]:

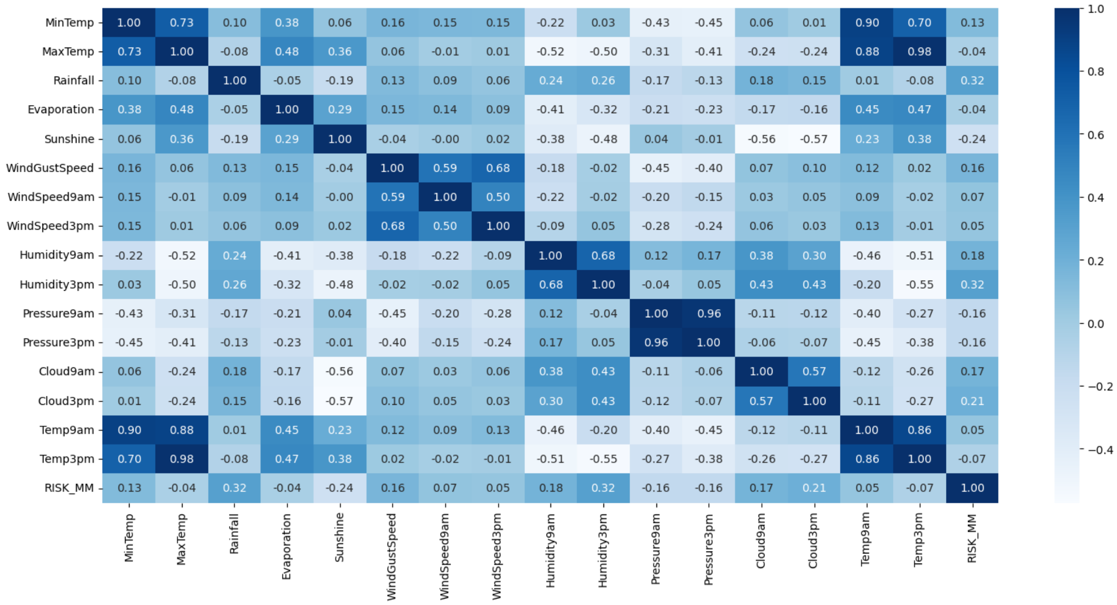
```
df.hist(figsize=(15,18), color='green',edgecolor='red',bins=20)
plt.show()
```



[43]:
```
plt.figure(figsize=(18,8))
sns.heatmap(df.corr(), annot=True, cmap='Blues',fmt=".2f")
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_10968\1072939161.py:2:
FutureWarning: The default value of numeric_only in DataFrame.corr is
deprecated. In a future version, it will default to False. Select only valid
columns or specify the value of numeric_only to silence this warning.
  sns.heatmap(df.corr(), annot=True, cmap='Blues',fmt=".2f")
```

```
[43]: <Axes: >
```



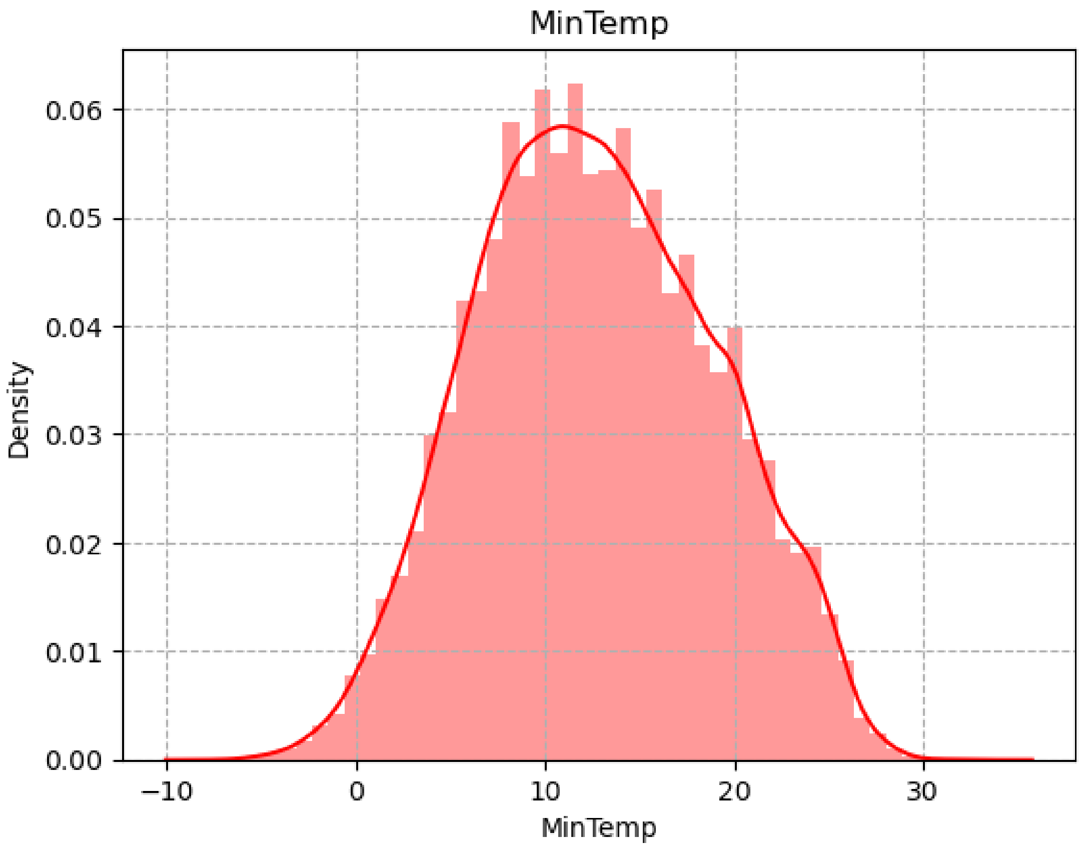
```
[40]: df.columns
```

```
[40]: Index(['Date', 'Location', 'MinTemp', 'MaxTemp', 'Rainfall', 'Evaporation',
             'Sunshine', 'WindGustDir', 'WindGustSpeed', 'WindDir9am', 'WindDir3pm',
             'WindSpeed9am', 'WindSpeed3pm', 'Humidity9am', 'Humidity3pm',
             'Pressure9am', 'Pressure3pm', 'Cloud9am', 'Cloud3pm', 'Temp9am',
             'Temp3pm', 'RainToday', 'RISK_MM', 'RainTomorrow'],
            dtype='object')
```

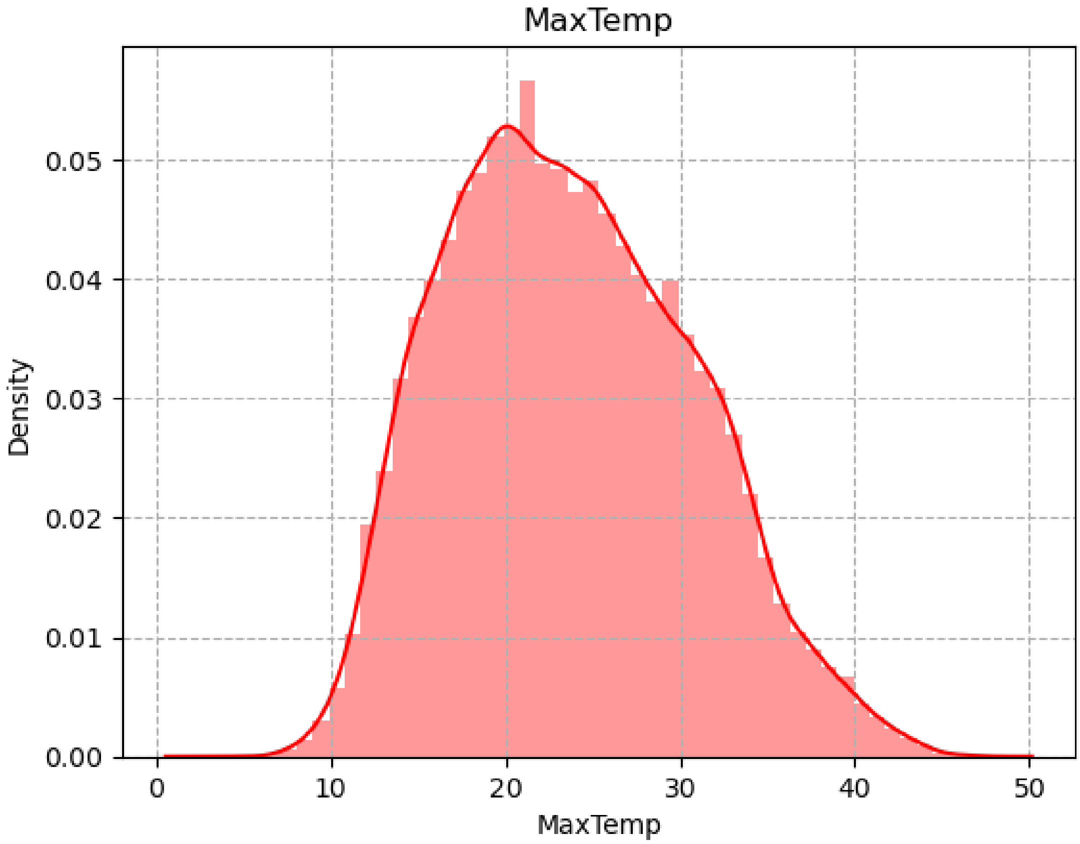
```
[51]: for x in continuous_feature:
          sns.distplot(df[x], color='red')
          plt.xlabel(x)
          plt.title(x)
          plt.grid(linestyle='--')
          plt.figure(figsize=(15,15))
          plt.show()
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_10968\4047360785.py:2: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
```

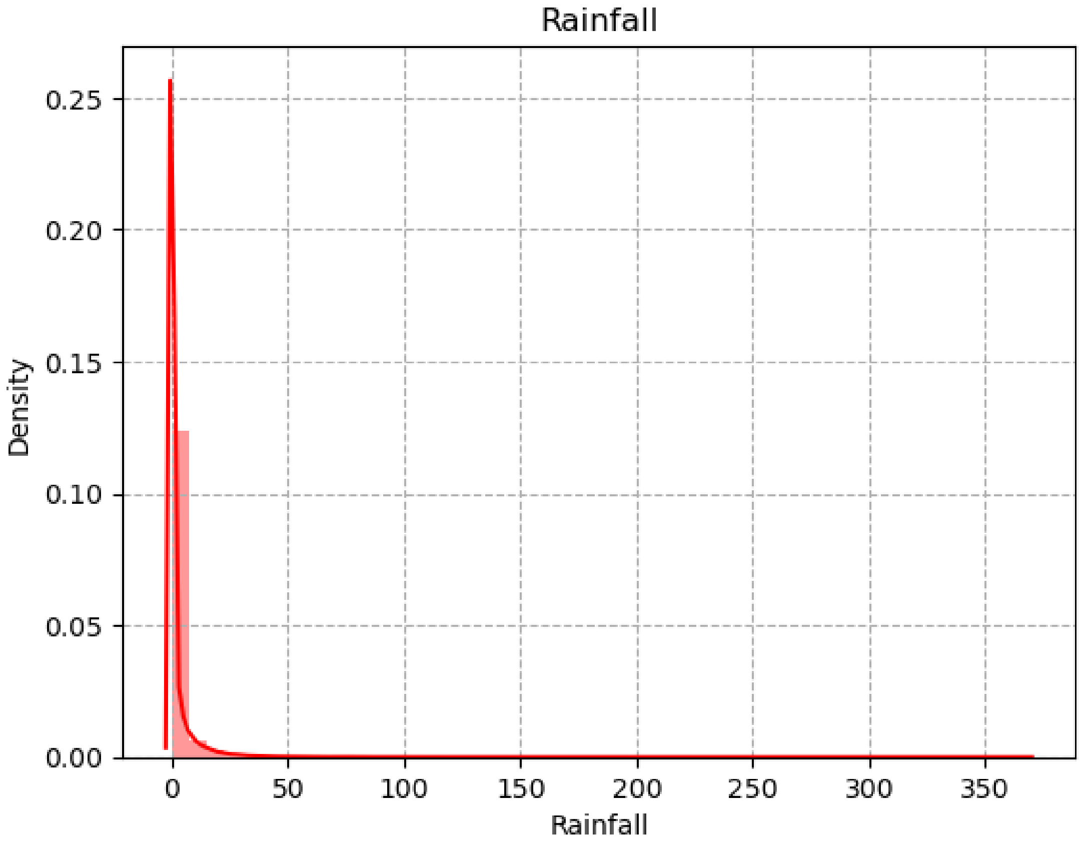
```
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df[x], color='red')
```



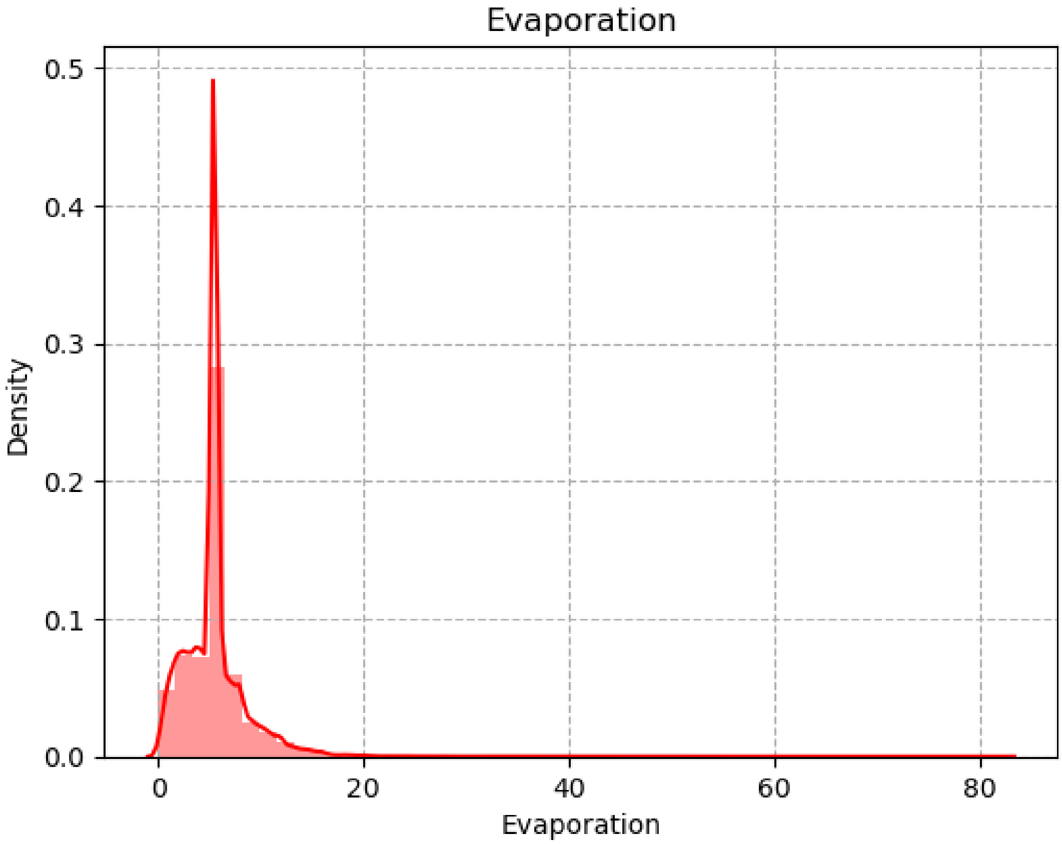
```
<Figure size 1500x1500 with 0 Axes>

C:\Users\Admin\AppData\Local\Temp\ipykernel_10968\4047360785.py:2: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df[x], color='red')
```



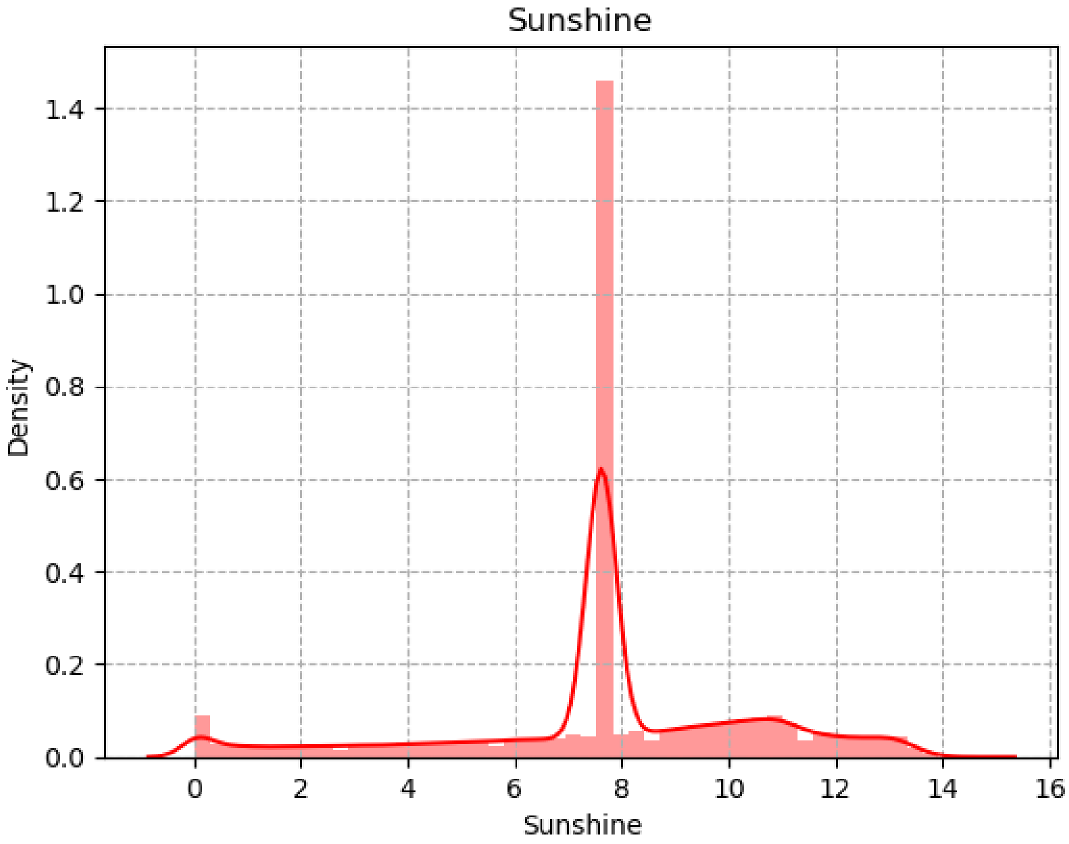
```
<Figure size 1500x1500 with 0 Axes>

C:\Users\Admin\AppData\Local\Temp\ipykernel_10968\4047360785.py:2: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df[x], color='red')
```



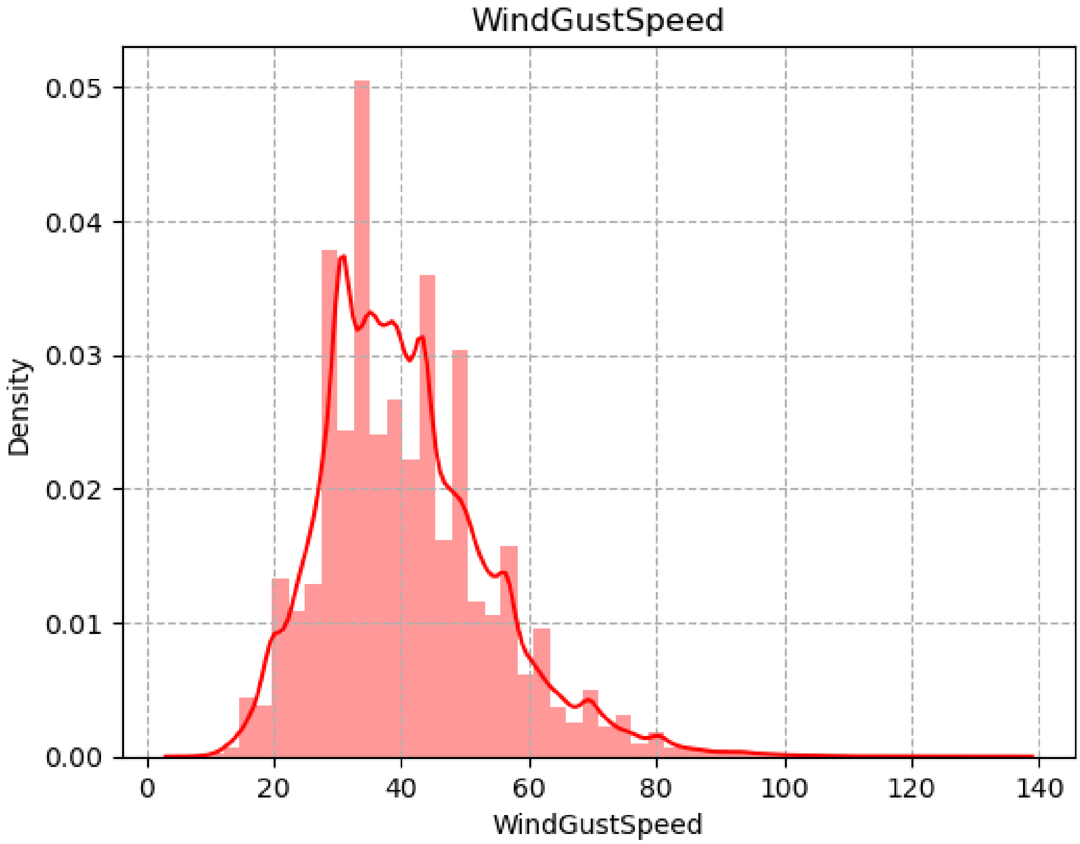
```
<Figure size 1500x1500 with 0 Axes>

C:\Users\Admin\AppData\Local\Temp\ipykernel_10968\4047360785.py:2: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df[x], color='red')
```



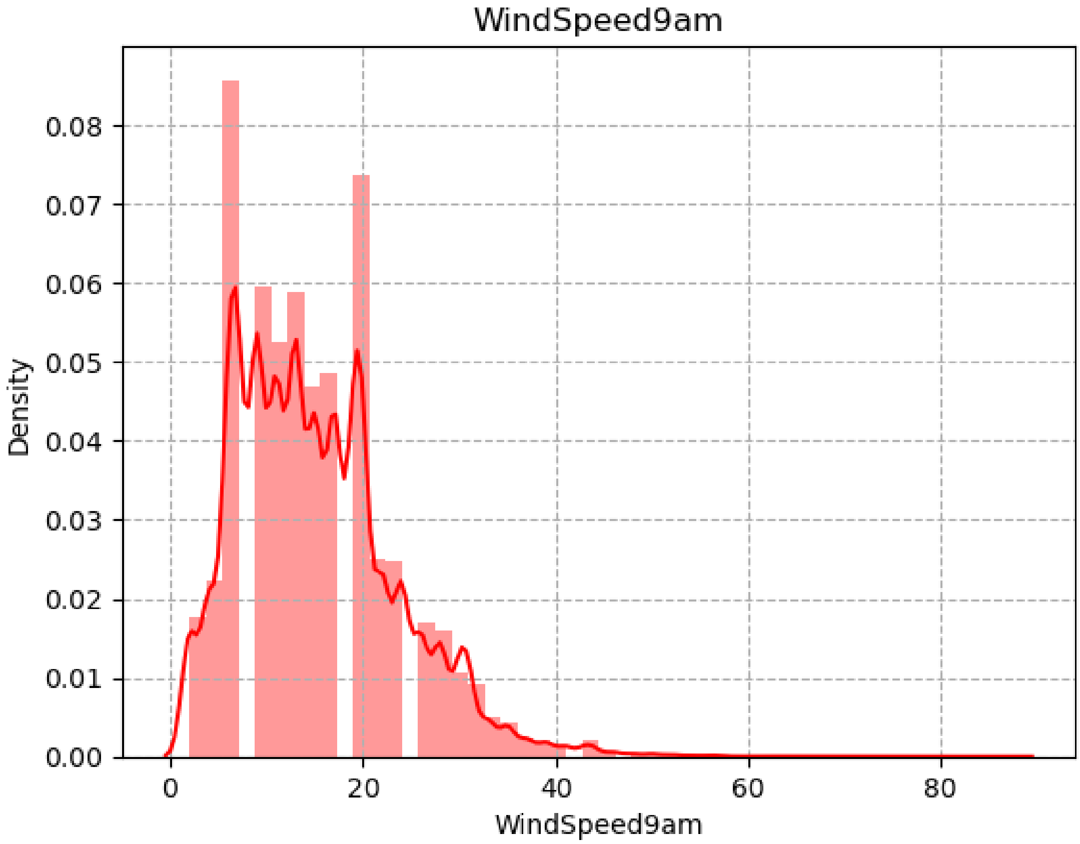
```
<Figure size 1500x1500 with 0 Axes>

C:\Users\Admin\AppData\Local\Temp\ipykernel_10968\4047360785.py:2: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df[x], color='red')
```



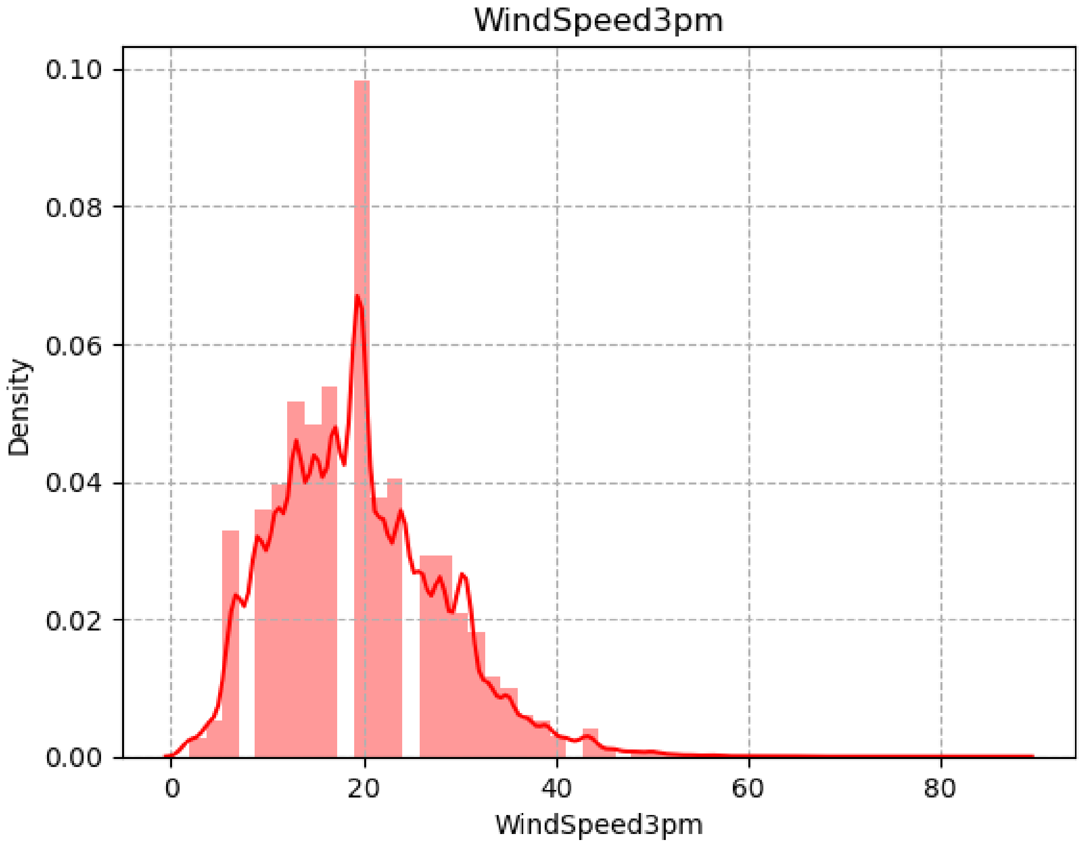
```
<Figure size 1500x1500 with 0 Axes>

C:\Users\Admin\AppData\Local\Temp\ipykernel_10968\4047360785.py:2: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df[x], color='red')
```



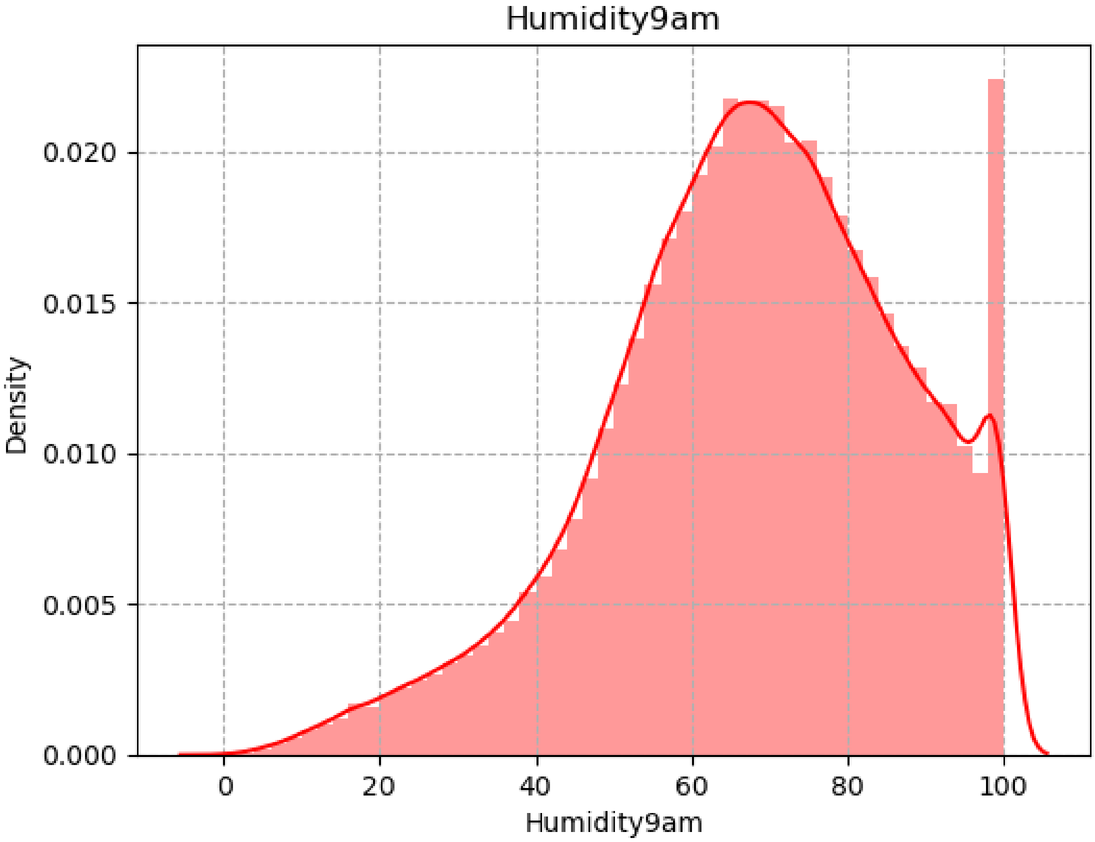
```
<Figure size 1500x1500 with 0 Axes>

C:\Users\Admin\AppData\Local\Temp\ipykernel_10968\4047360785.py:2: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df[x], color='red')
```



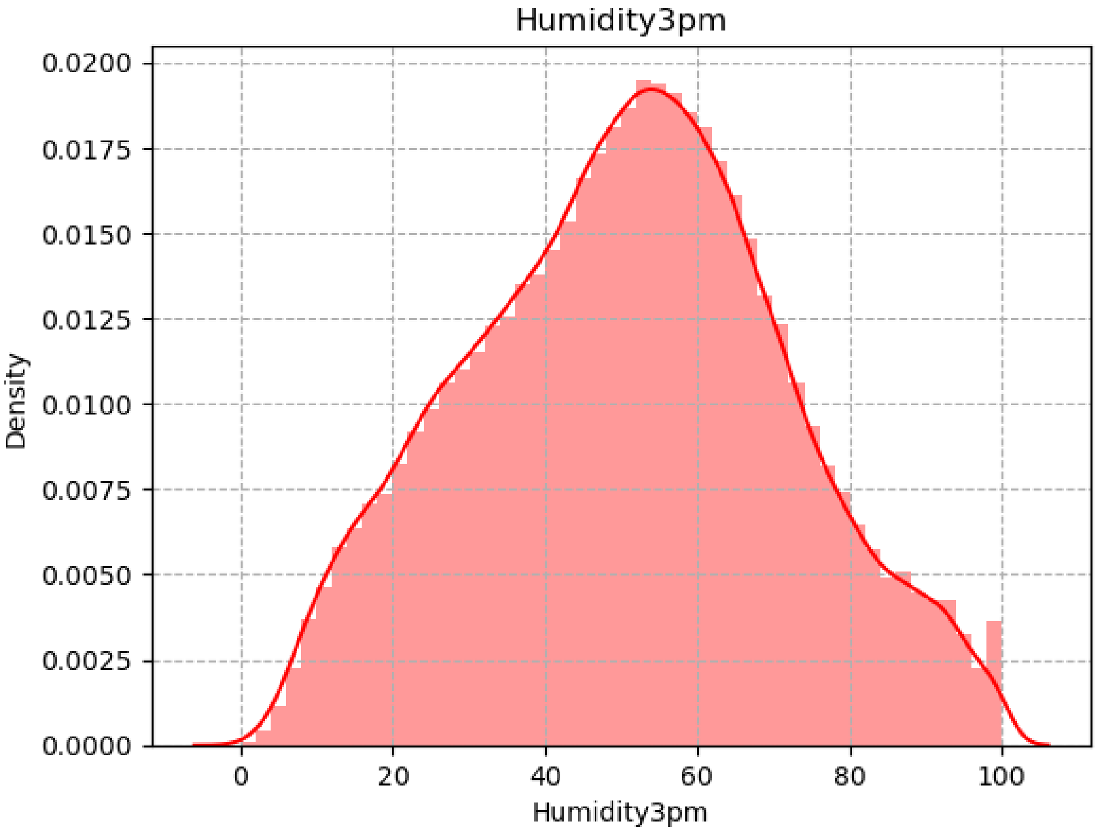
```
<Figure size 1500x1500 with 0 Axes>

C:\Users\Admin\AppData\Local\Temp\ipykernel_10968\4047360785.py:2: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df[x], color='red')
```



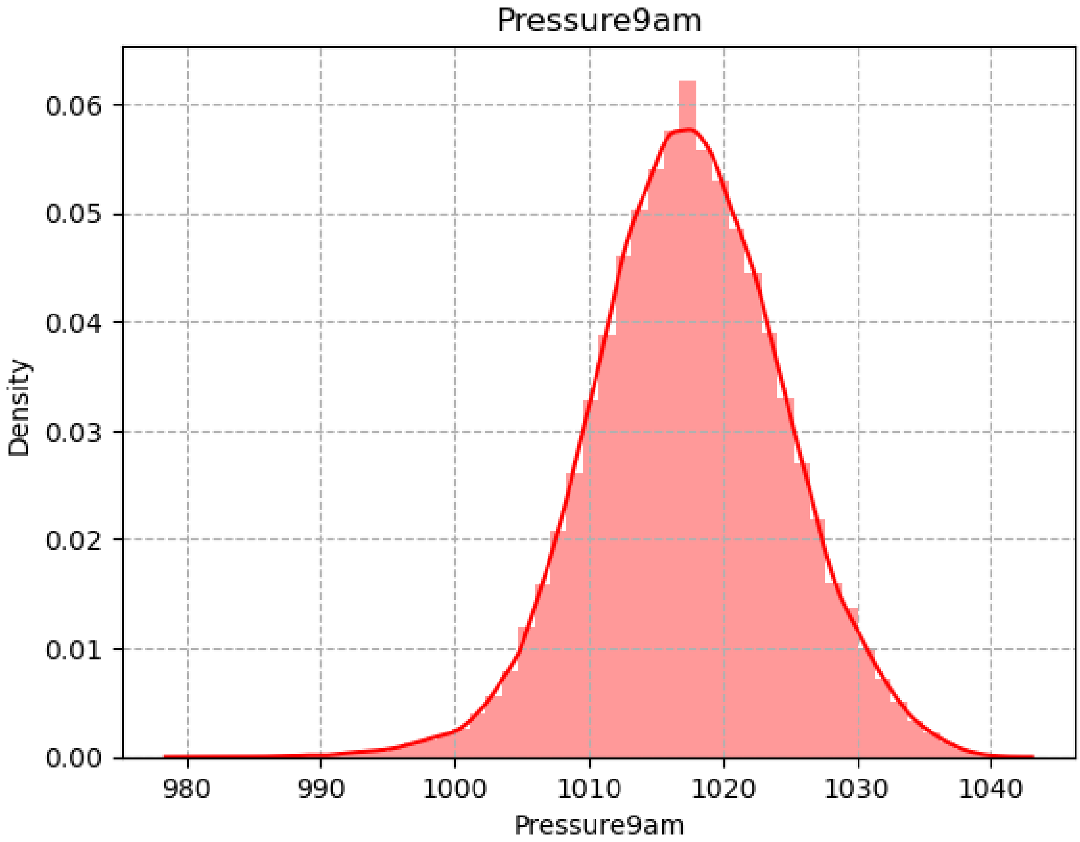
```
<Figure size 1500x1500 with 0 Axes>

C:\Users\Admin\AppData\Local\Temp\ipykernel_10968\4047360785.py:2: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df[x], color='red')
```



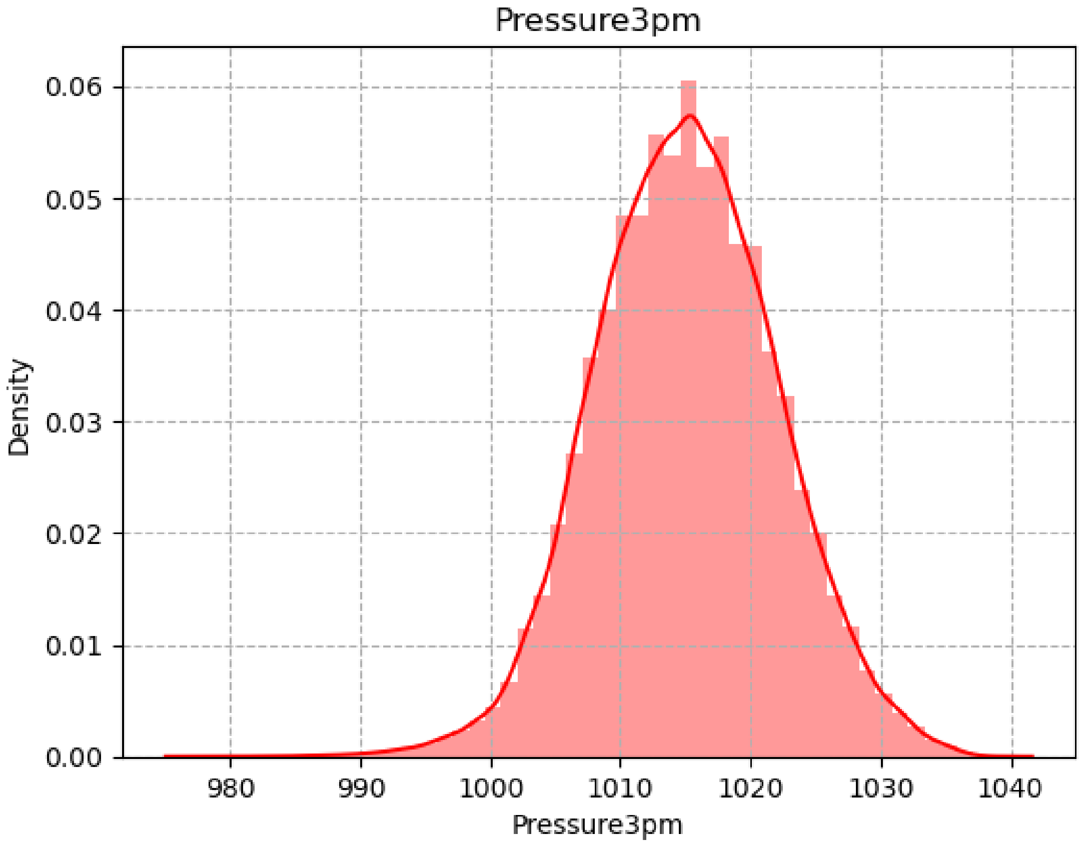
```
<Figure size 1500x1500 with 0 Axes>

C:\Users\Admin\AppData\Local\Temp\ipykernel_10968\4047360785.py:2: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df[x], color='red')
```



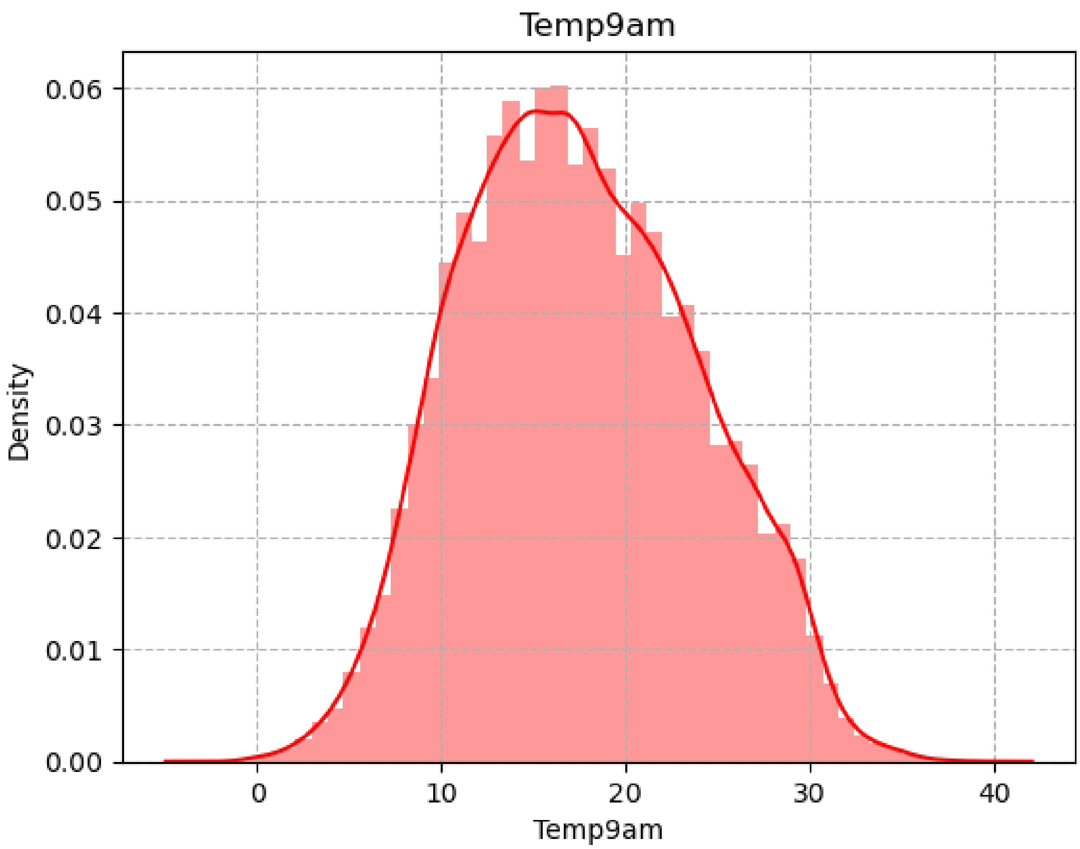
```
<Figure size 1500x1500 with 0 Axes>
C:\Users\Admin\AppData\Local\Temp\ipykernel_10968\4047360785.py:2: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df[x], color='red')
```



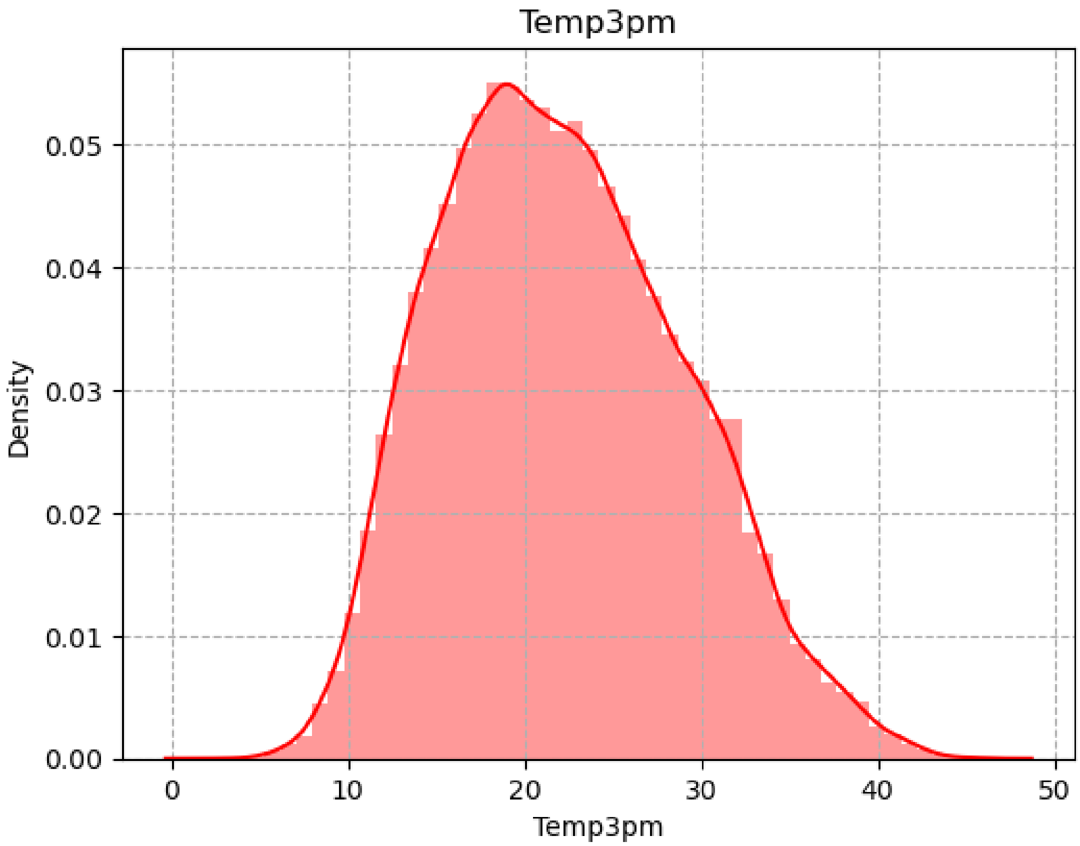
```
<Figure size 1500x1500 with 0 Axes>

C:\Users\Admin\AppData\Local\Temp\ipykernel_10968\4047360785.py:2: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df[x], color='red')
```



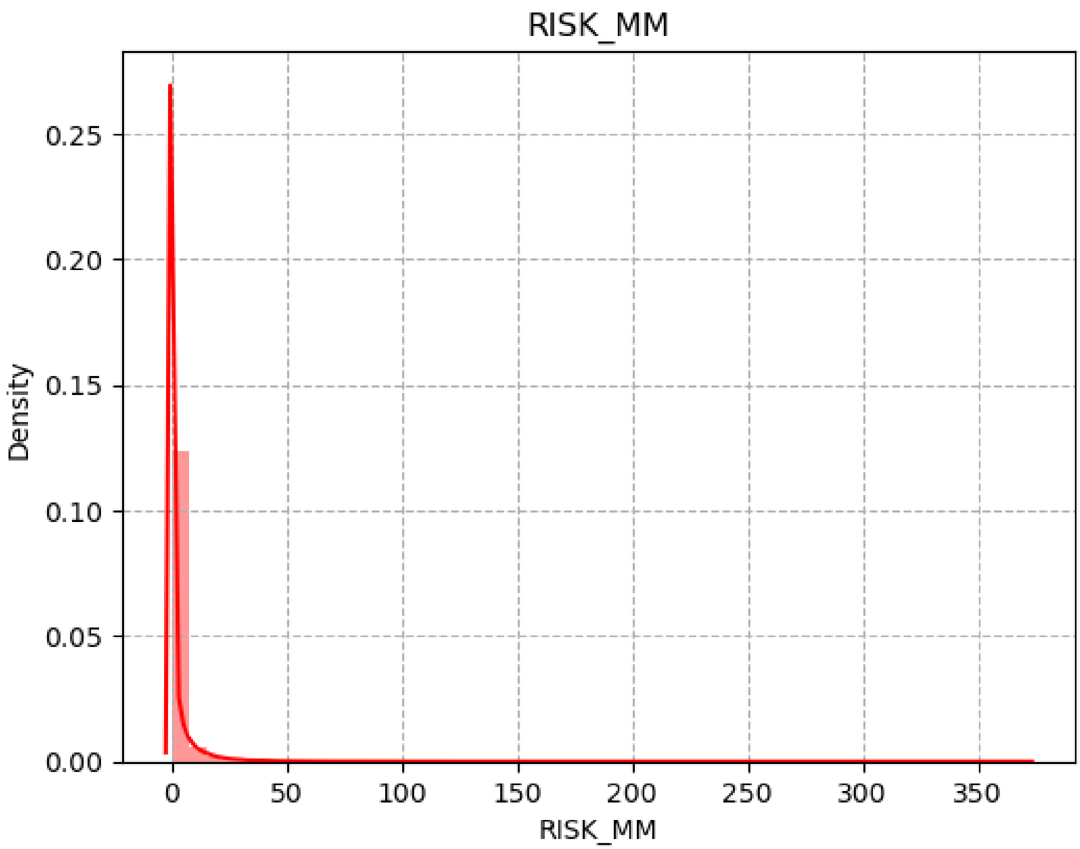
```
<Figure size 1500x1500 with 0 Axes>

C:\Users\Admin\AppData\Local\Temp\ipykernel_10968\4047360785.py:2: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df[x], color='red')
```



```
<Figure size 1500x1500 with 0 Axes>

C:\Users\Admin\AppData\Local\Temp\ipykernel_10968\4047360785.py:2: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df[x], color='red')
```

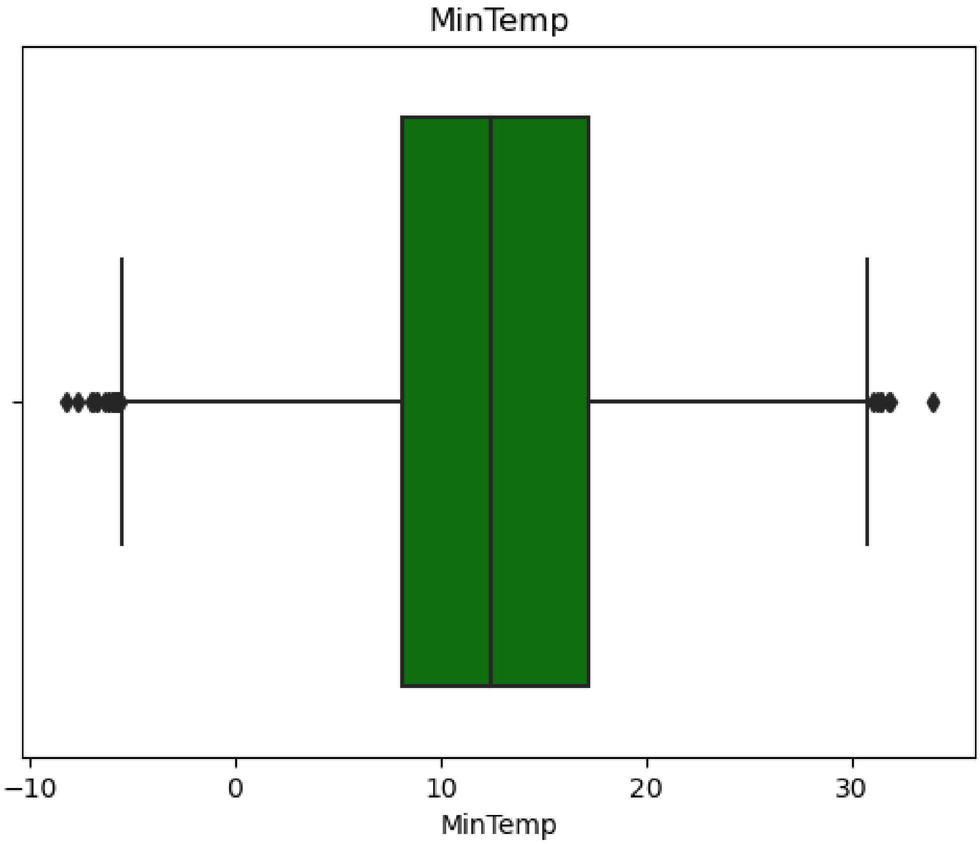


```
<Figure size 1500x1500 with 0 Axes>
C:\Users\Admin\AppData\Local\Temp\ipykernel_10968\4047360785.py:2: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df[x], color='red')
```

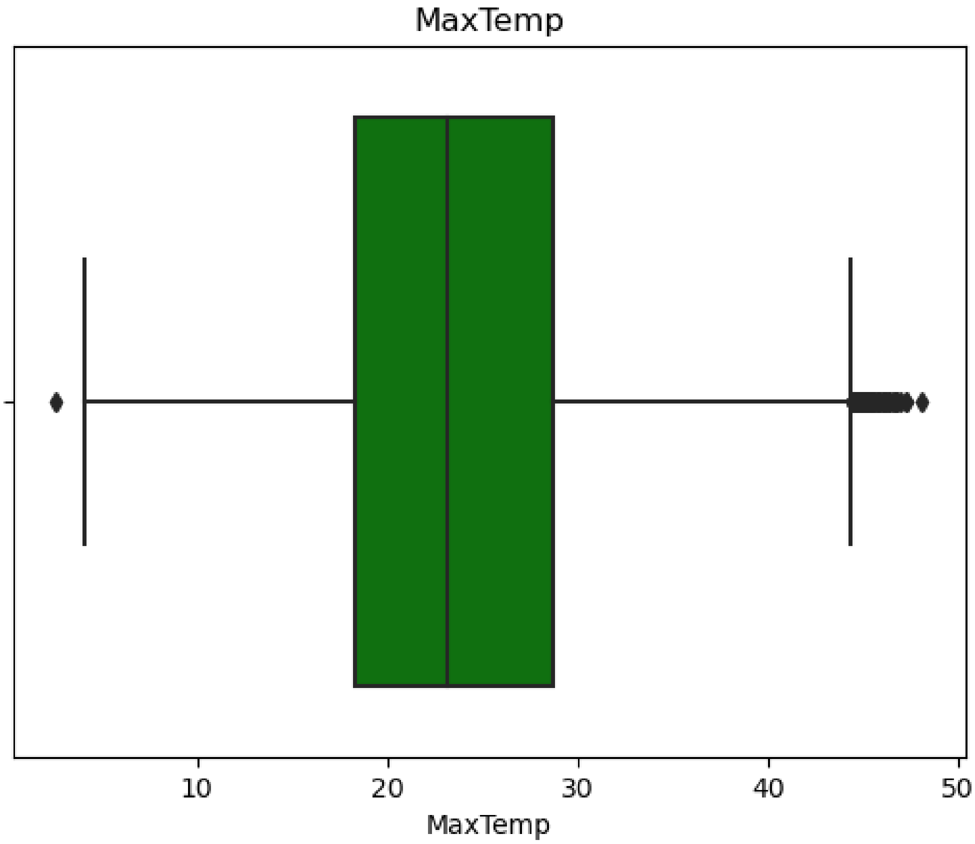


```
<Figure size 1500x1500 with 0 Axes>
```

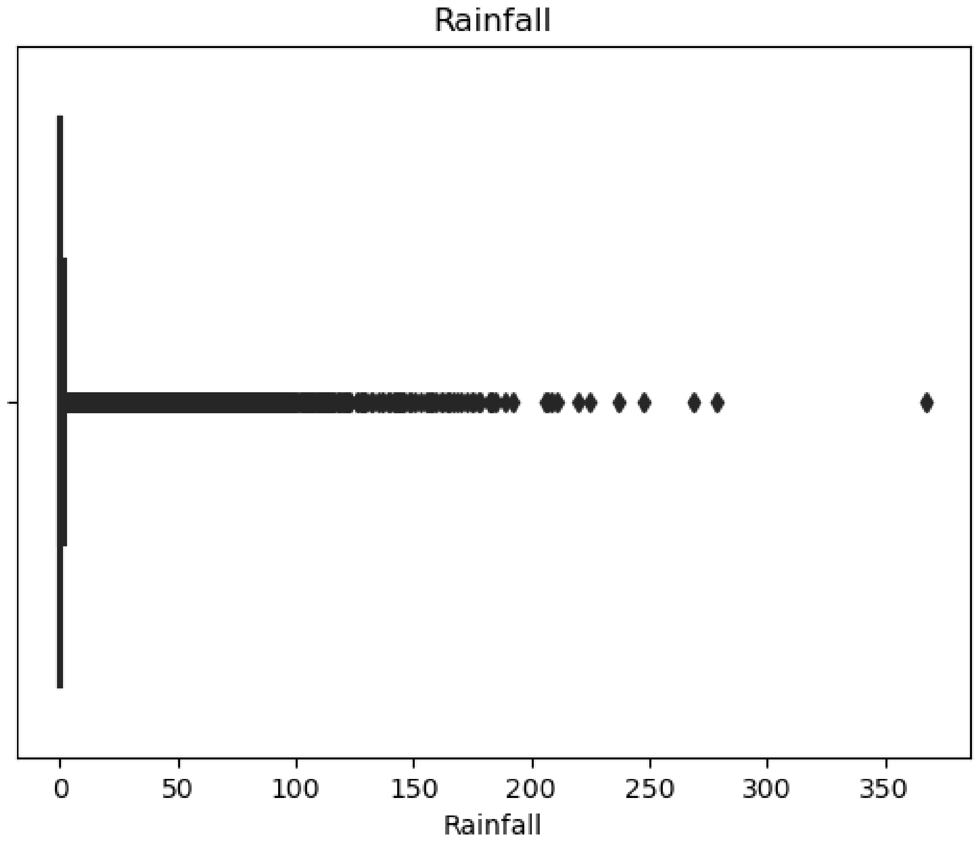
```
[53]: for feature in continuous_feature:
          sns.boxplot(x = df[feature], color='green')
          plt.title(feature)
          plt.figure(figsize=(8,8))
          plt.show()
```


MinTemp
−10
0
10
20
30
MinTemp

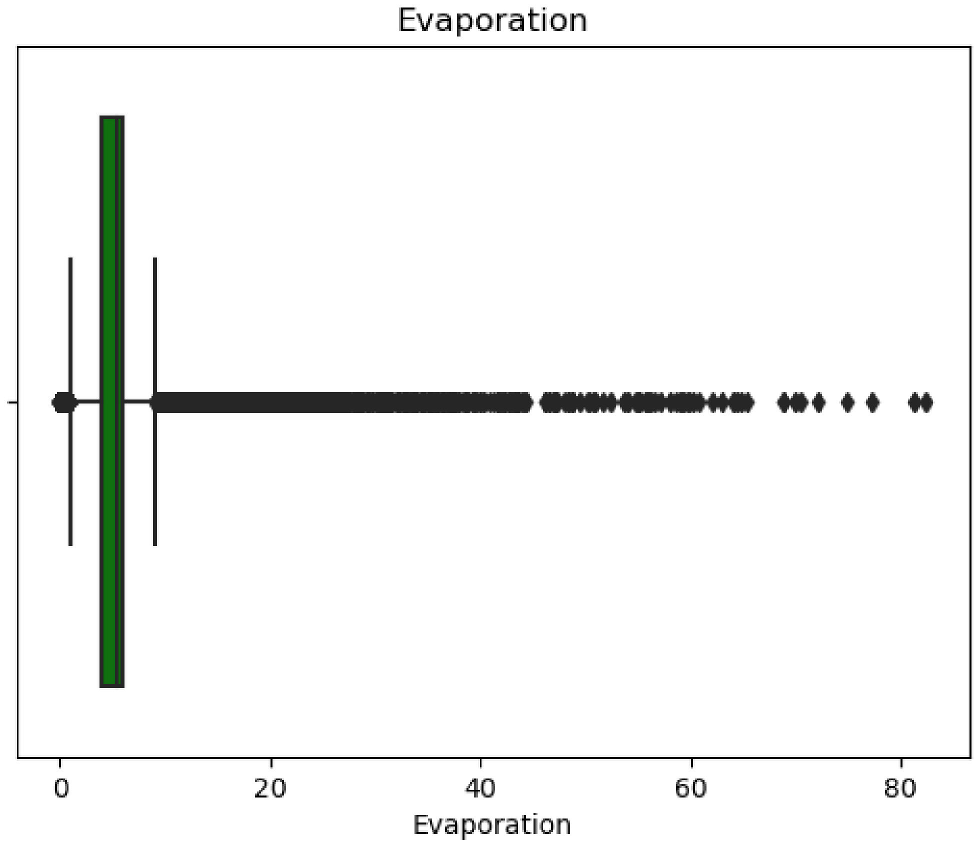
```
<Figure size 800x800 with 0 Axes>
```


MaxTemp
10
20
30
40
50
MaxTemp

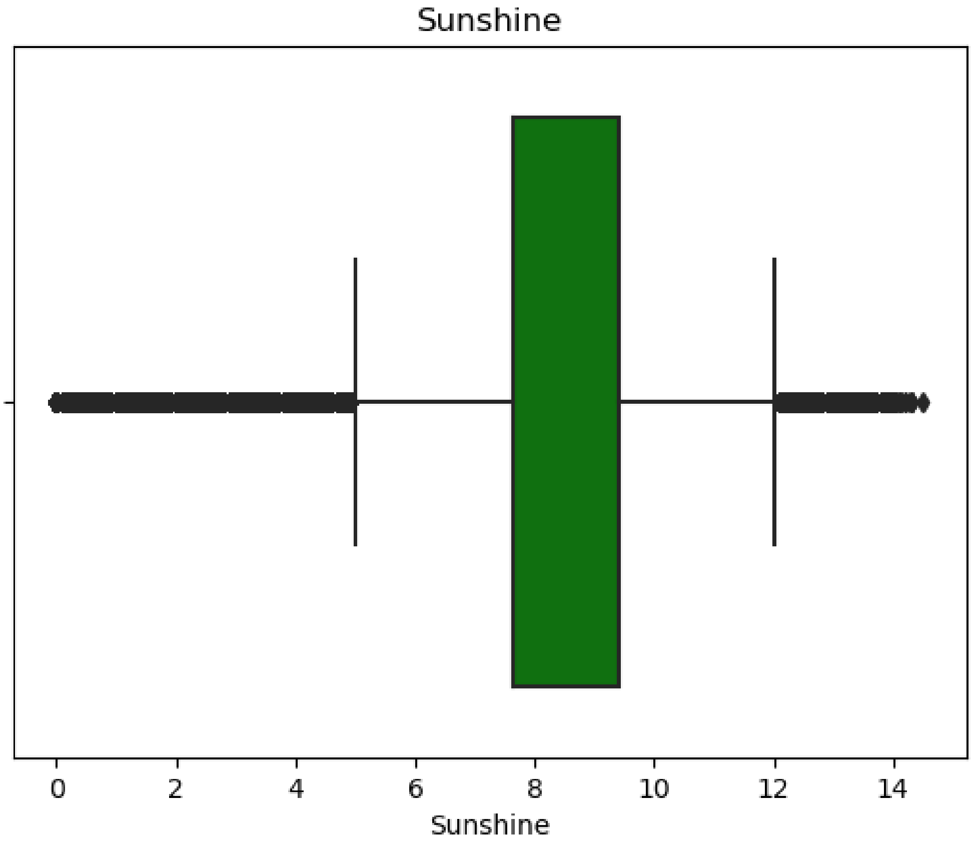
```
<Figure size 800x800 with 0 Axes>
```


Rainfall
0
50
100
150
200
250
300
350
Rainfall

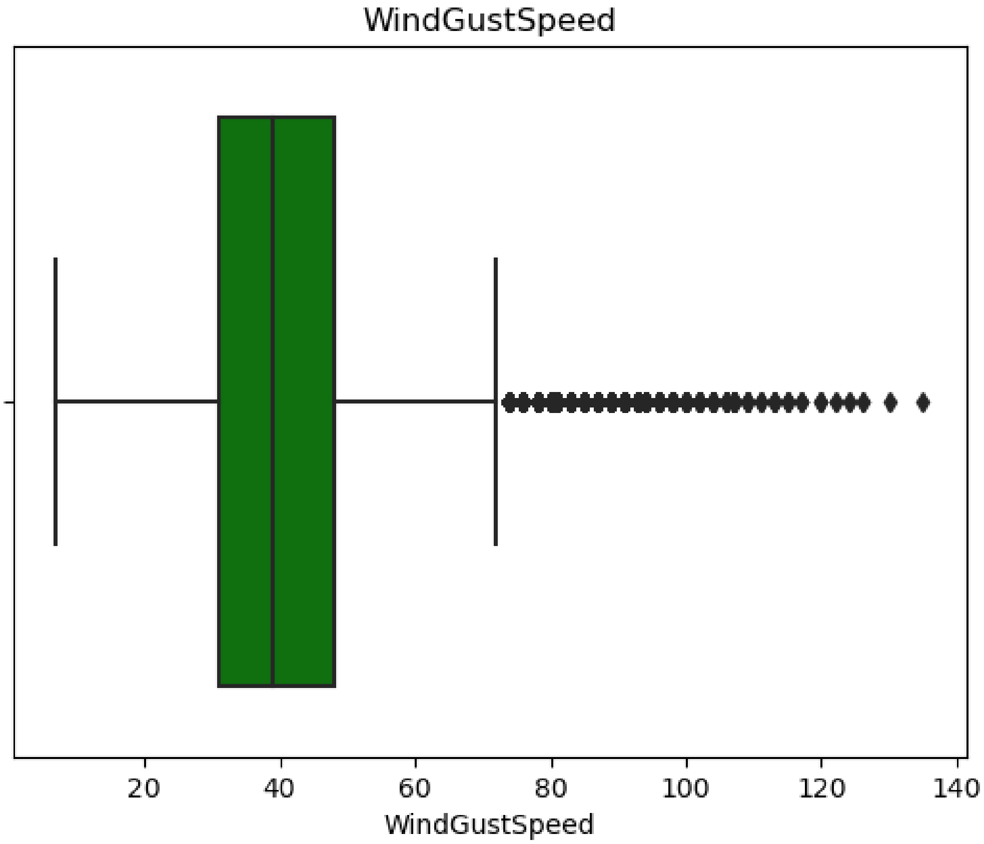
```
<Figure size 800x800 with 0 Axes>
```


Evaporation
0
20
40
60
80
Evaporation

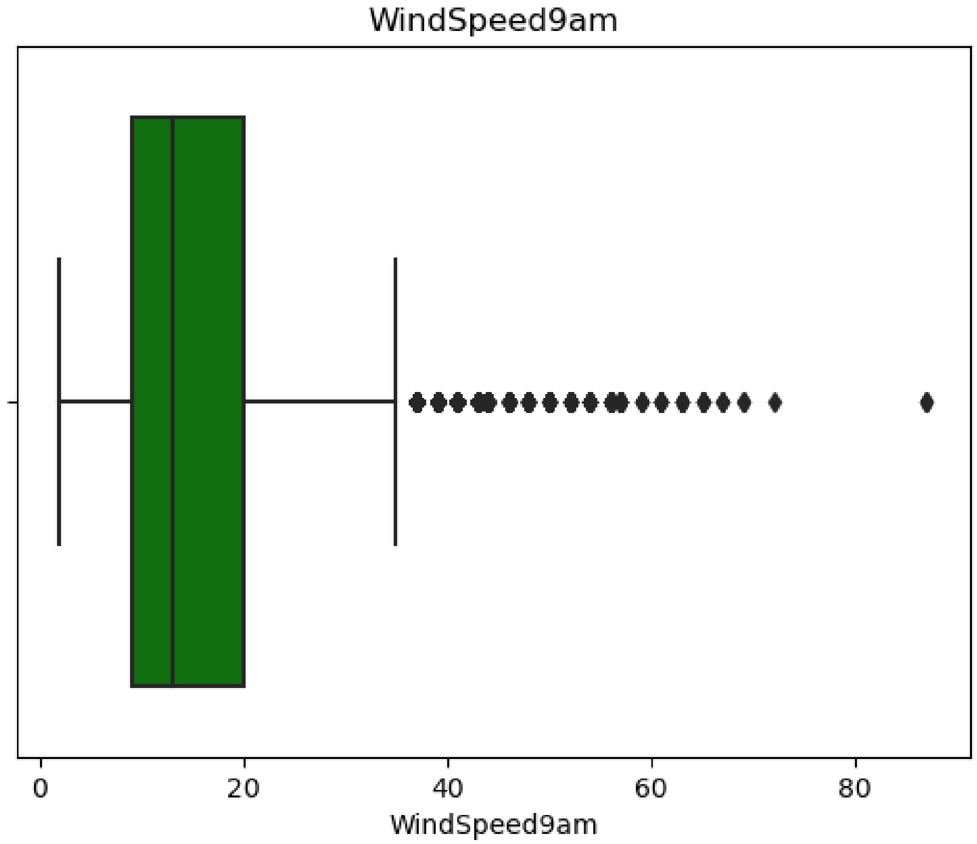
```
<Figure size 800x800 with 0 Axes>
```


Sunshine
0
2
4
6
8
10
12
14
Sunshine

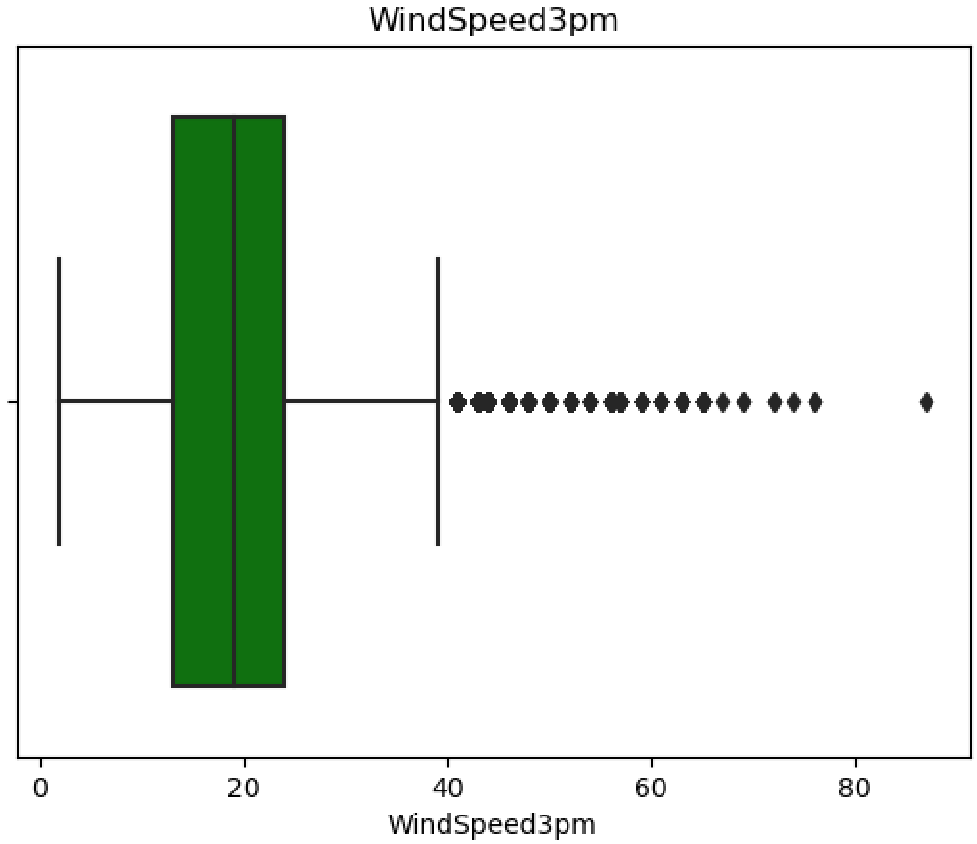
```
<Figure size 800x800 with 0 Axes>
```


WindGustSpeed
20
40
60
80
100
120
140
WindGustSpeed

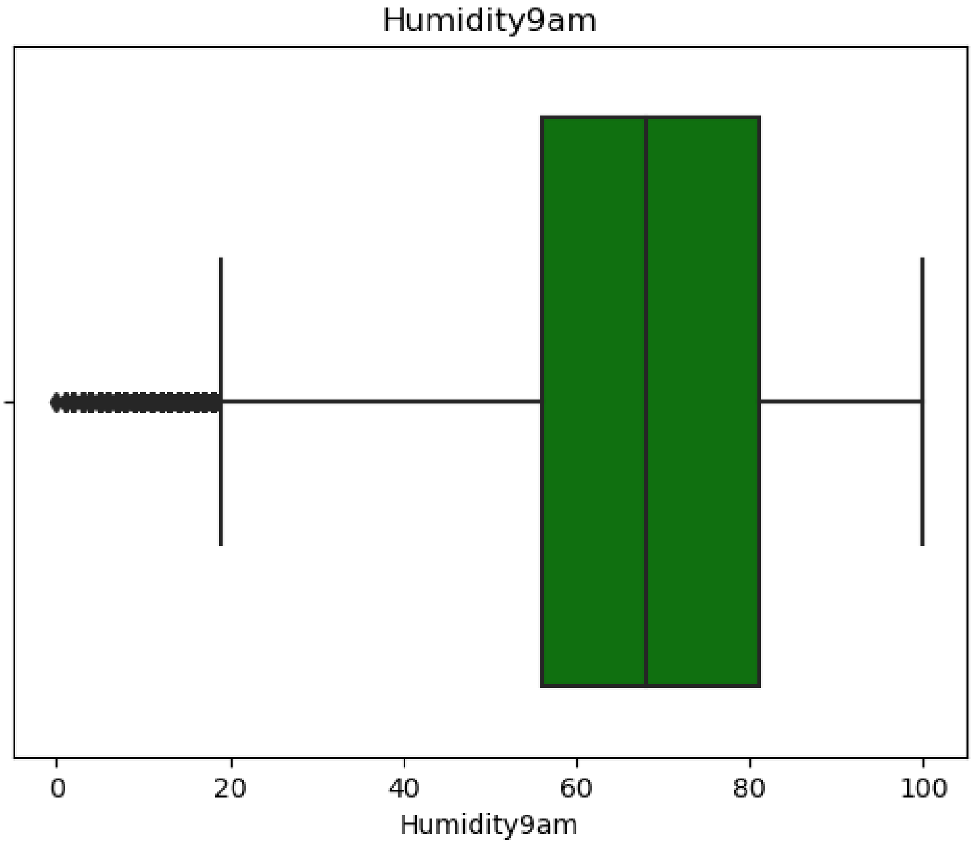
```
<Figure size 800x800 with 0 Axes>
```


WindSpeed9am
0
20
40
60
80
WindSpeed9am

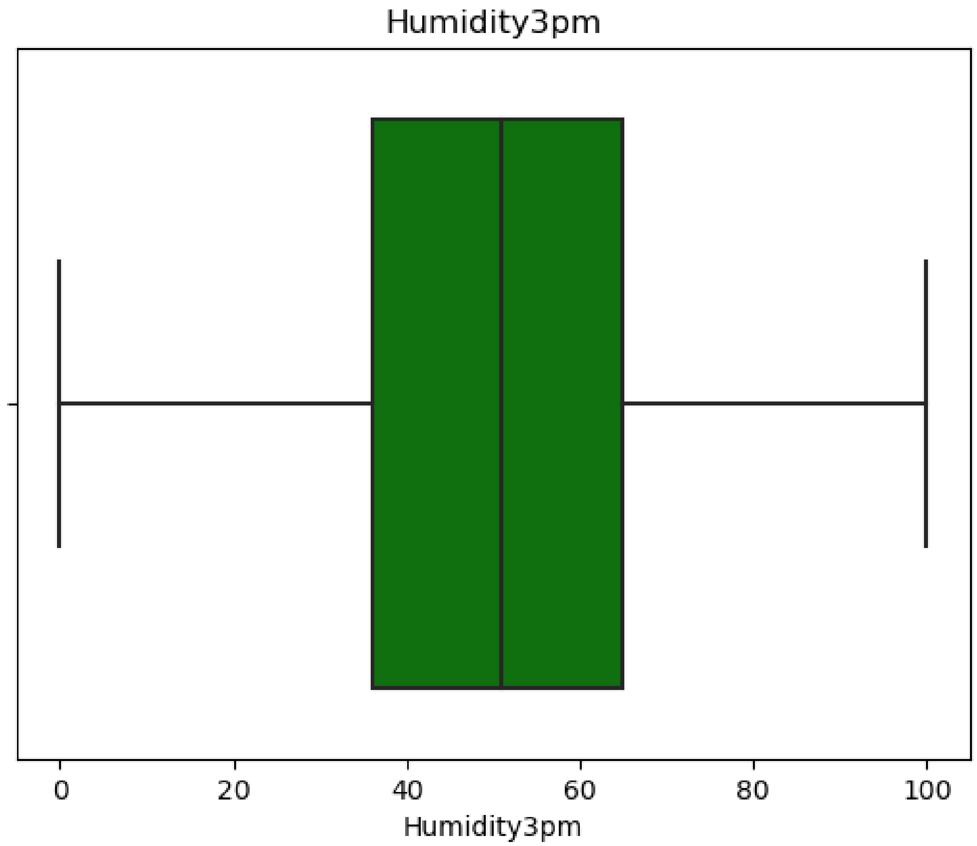
```
<Figure size 800x800 with 0 Axes>
```


WindSpeed3pm
0
20
40
60
80
WindSpeed3pm

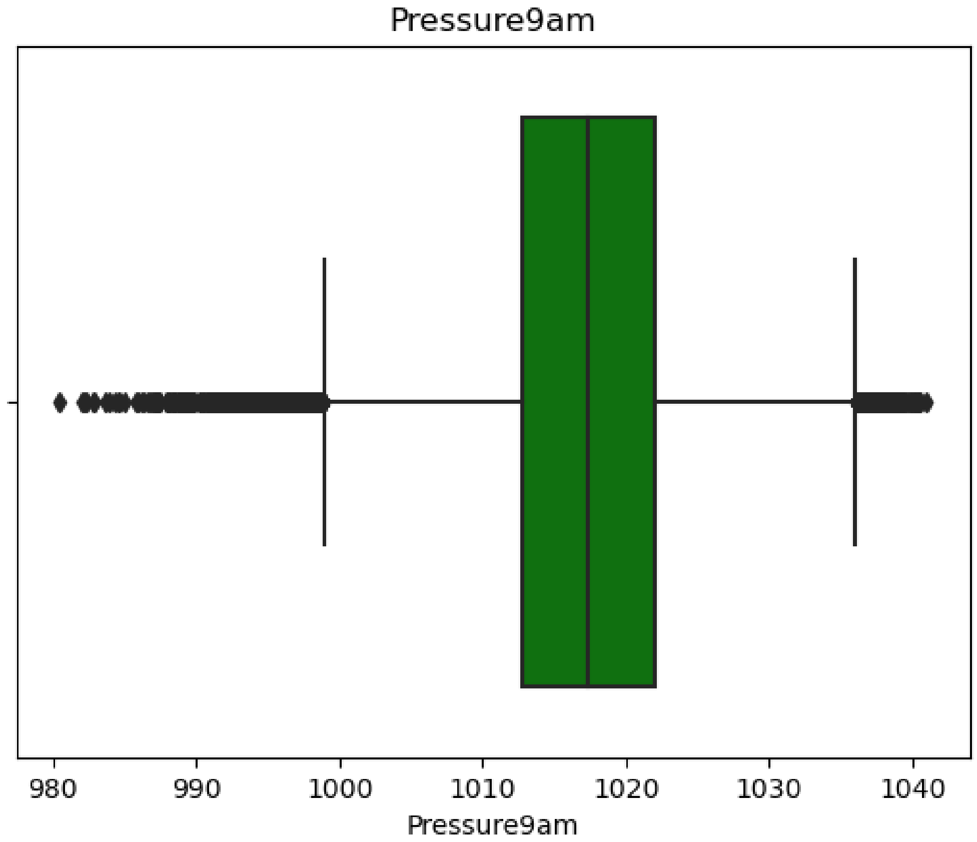
```
<Figure size 800x800 with 0 Axes>
```


Humidity9am
0
20
40
60
80
100
Humidity9am

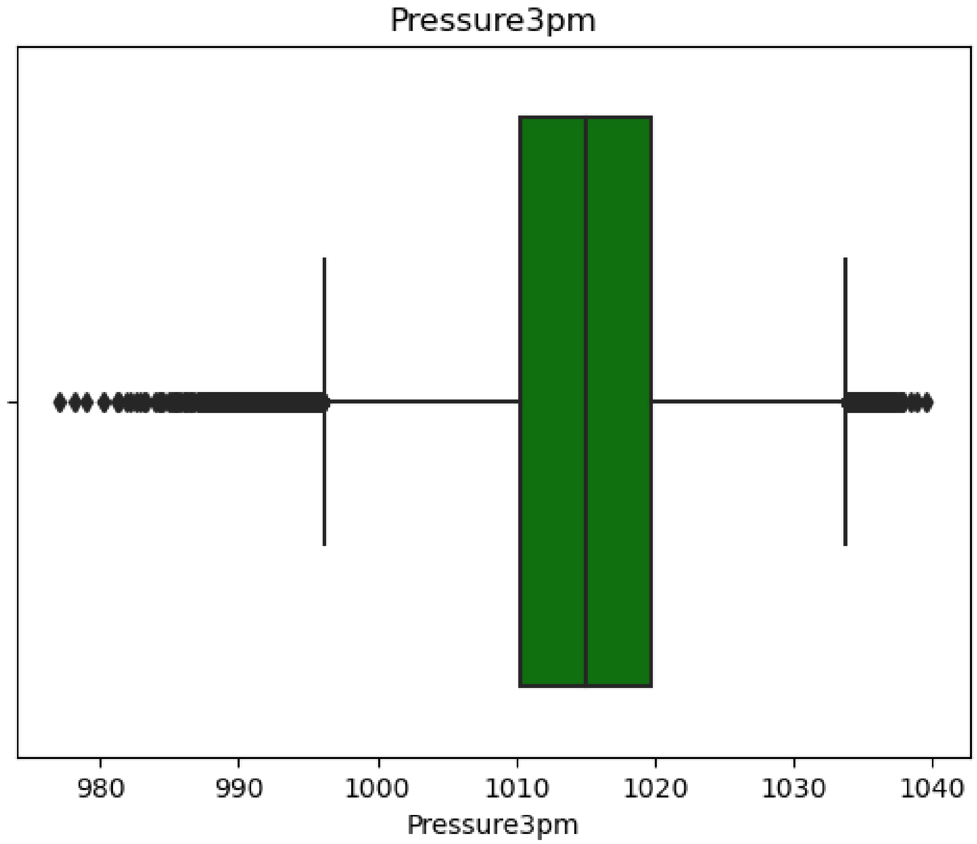
```
<Figure size 800x800 with 0 Axes>
```


Humidity3pm
0
20
40
60
80
100
Humidity3pm

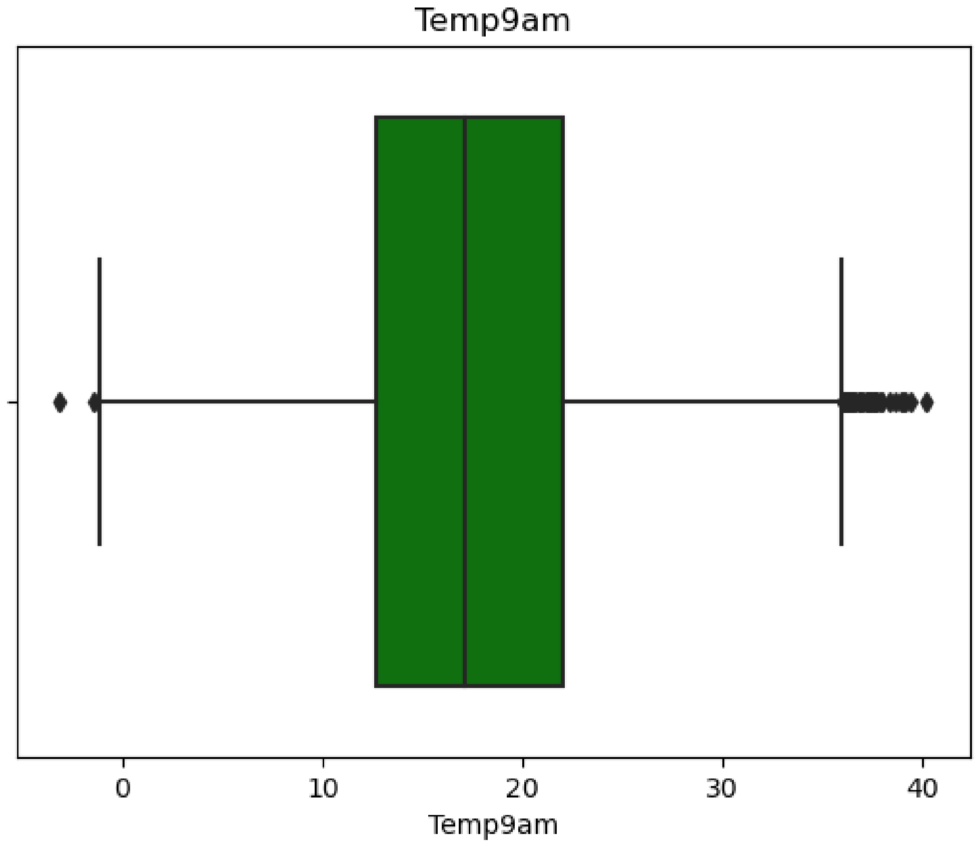
```
<Figure size 800x800 with 0 Axes>
```


Pressure9am
980
990
1000
1010
1020
1030
1040
Pressure9am

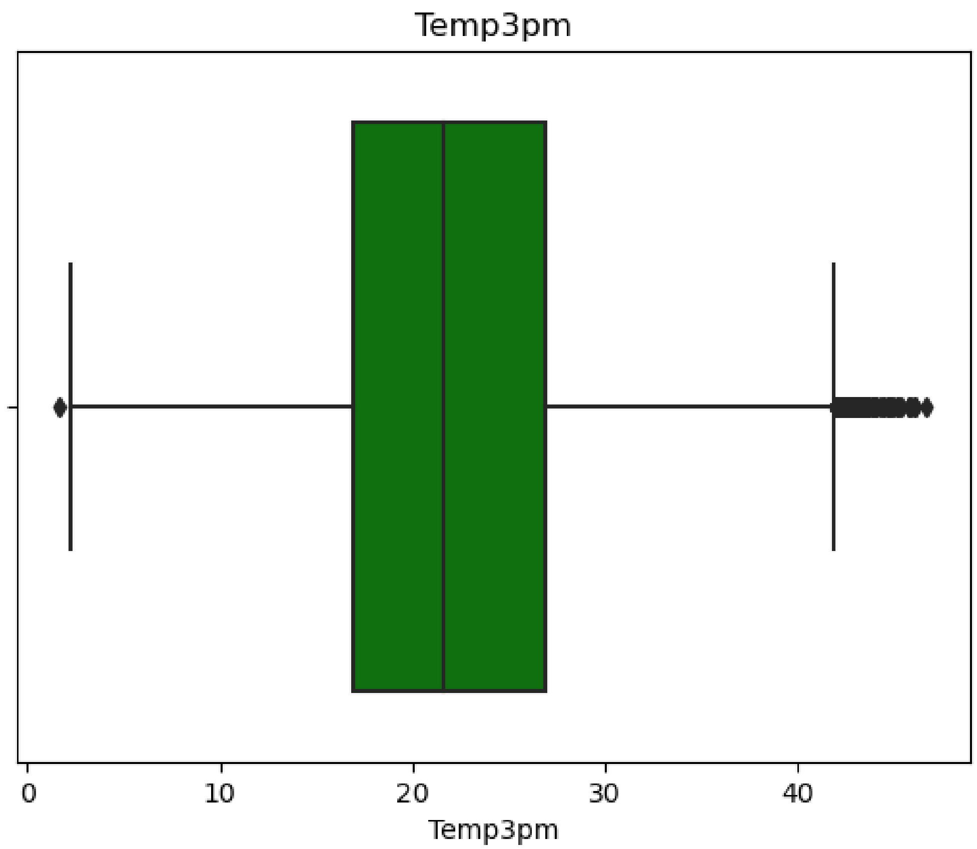
```
<Figure size 800x800 with 0 Axes>
```


Pressure3pm
980
990
1000
1010
1020
1030
1040
Pressure3pm

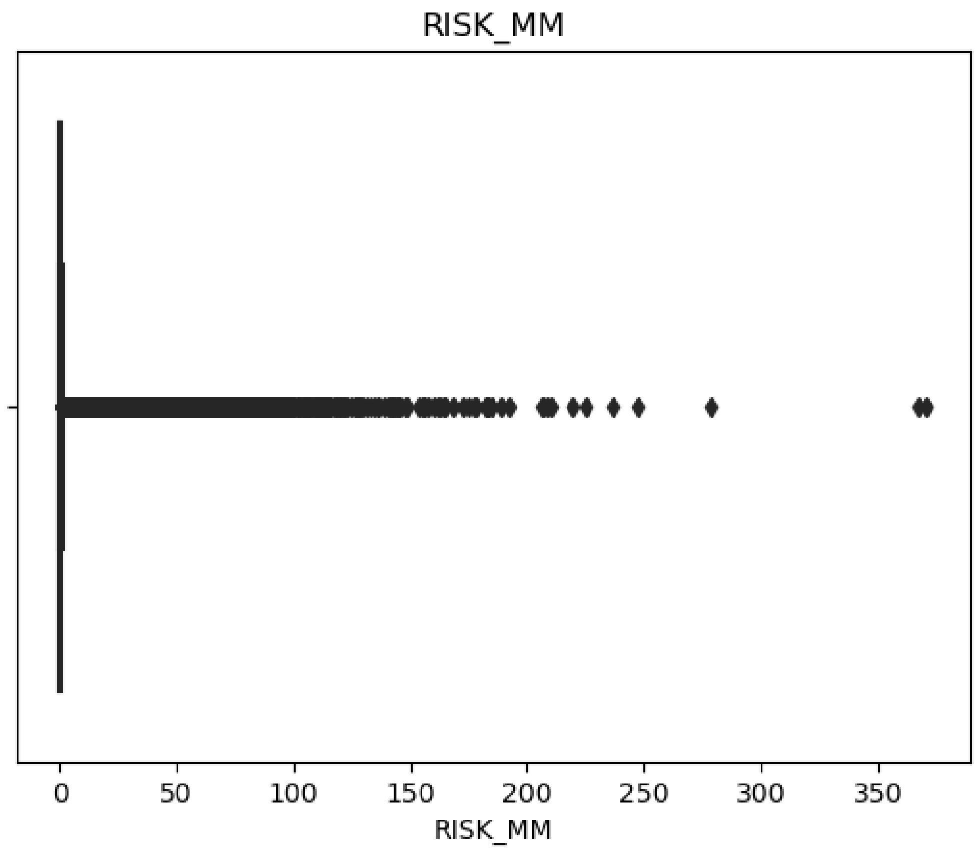
```
<Figure size 800x800 with 0 Axes>
```


Temp9am
0
10
20
30
40
Temp9am

```
<Figure size 800x800 with 0 Axes>
```


Temp3pm
0
10
20
30
40
Temp3pm

```
<Figure size 800x800 with 0 Axes>
```



```
<Figure size 800x800 with 0 Axes>
```

[58]:
```
df['RainToday'] = pd.get_dummies(df['RainToday'], drop_first=True)
df['RainTomorrow'] = pd.get_dummies(df['RainTomorrow'], drop_first=True)
df.head()
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_10968\574245326.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df['RainToday'] = pd.get_dummies(df['RainToday'], drop_first=True)
C:\Users\Admin\AppData\Local\Temp\ipykernel_10968\574245326.py:2:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
```

```
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df['RainTomorrow'] = pd.get_dummies(df['RainTomorrow'], drop_first=True)
```

[58]:

```
         Date Location  MinTemp  MaxTemp  Rainfall  Evaporation  Sunshine  \
0  2008-12-01   Albury     13.4     22.9       0.6     5.469824  7.624853
1  2008-12-02   Albury      7.4     25.1       0.0     5.469824  7.624853
2  2008-12-03   Albury     12.9     25.7       0.0     5.469824  7.624853
3  2008-12-04   Albury      9.2     28.0       0.0     5.469824  7.624853
4  2008-12-05   Albury     17.5     32.3       1.0     5.469824  7.624853

  WindGustDir  WindGustSpeed WindDir9am WindDir3pm  WindSpeed9am  \
0           W           44.0          W        WNW          20.0
1         WNW           44.0        NNW        WSW           4.0
2         WSW           46.0          W        WSW          19.0
3          NE           24.0         SE          E          11.0
4           W           41.0        ENE         NW           7.0

   WindSpeed3pm  Humidity9am  Humidity3pm  Pressure9am  Pressure3pm  Cloud9am  \
0          24.0         71.0         22.0       1007.7       1007.1  8.000000
1          22.0         44.0         25.0       1010.6       1007.8  4.437189
2          26.0         38.0         30.0       1007.6       1008.7  4.437189
3           9.0         45.0         16.0       1017.6       1012.8  4.437189
4          20.0         82.0         33.0       1010.8       1006.0  7.000000

   Cloud3pm  Temp9am  Temp3pm  RainToday  RISK_MM  RainTomorrow
0  4.503167     16.9     21.8          0      0.0             0
1  4.503167     17.2     24.3          0      0.0             0
2  2.000000     21.0     23.2          0      0.0             0
3  4.503167     18.1     26.5          0      1.0             0
4  8.000000     17.8     29.7          0      0.2             0
```

[59]:

```
df1 = df.groupby(['Location'])['RainTomorrow'].value_counts().sort_values().
  ↪unstack()
```

[60]:

```
df1
```

[60]:

```
RainTomorrow         0    1
Location
Adelaide          2115  625
Albury            1913  527
AliceSprings      2517  227
BadgerysCreek     1869  465
Ballarat          2109  745
Bendigo           2198  515
Brisbane          2358  662
Cairns            1989  910
Canberra          2222  503
```

```
Cobar             2445   359
CoffsHarbour      1781   748
Dartmoor          1524   770
Darwin            2300   817
GoldCoast         2088   733
Hobart            2350   739
Katherine          568   102
Launceston        1169   369
Melbourne         1712   521
MelbourneAirport  2296   638
Mildura           2582   315
Moree             2293   336
MountGambier      2010   876
Nhil              1282   236
NorahHead         2011   774
NorfolkIsland     1981   883
Nuriootpa         2240   550
PearceRAAF        2060   398
Perth             2419   618
PerthAirport      2367   556
Portland          1789  1031
Richmond          1624   424
Sale              2164   571
Sydney            1669   590
SydneyAirport     2182   747
Townsville        2393   491
Tuggeranong       1887   429
Uluru             1336   110
WaggaWagga        2292   508
Walpole           1638   864
Watsonia          2050   685
Williamtown       1683   512
Witchcliffe       1629   689
Wollongong        2109   658
Woomera           2693   193
```

[61]:

```
df1[1].sort_values(ascending=False)
```

[61]:

```
Location
Portland            1031
Cairns               910
NorfolkIsland        883
MountGambier         876
Walpole              864
Darwin               817
NorahHead            774
Dartmoor             770
```

```
CoffsHarbour        748
SydneyAirport       747
Ballarat            745
Hobart              739
GoldCoast           733
Witchcliffe         689
Watsonia            685
Brisbane            662
Wollongong          658
MelbourneAirport    638
Adelaide            625
Perth               618
Sydney              590
Sale                571
PerthAirport        556
Nuriootpa           550
Albury              527
Melbourne           521
Bendigo             515
Williamtown         512
WaggaWagga          508
Canberra            503
Townsville          491
BadgerysCreek       465
Tuggeranong         429
Richmond            424
PearceRAAF          398
Launceston          369
Cobar               359
Moree               336
Mildura             315
Nhil                236
AliceSprings        227
Woomera             193
Uluru               110
Katherine           102
Name: 1, dtype: int64
```

[66]:

```
df1[1].sort_values(ascending = False ).index
```

[66]:

```
Index(['Portland', 'Cairns', 'NorfolkIsland', 'MountGambier', 'Walpole',
       'Darwin', 'NorahHead', 'Dartmoor', 'CoffsHarbour', 'SydneyAirport',
       'Ballarat', 'Hobart', 'GoldCoast', 'Witchcliffe', 'Watsonia',
       'Brisbane', 'Wollongong', 'MelbourneAirport', 'Adelaide', 'Perth',
       'Sydney', 'Sale', 'PerthAirport', 'Nuriootpa', 'Albury', 'Melbourne',
       'Bendigo', 'Williamtown', 'WaggaWagga', 'Canberra', 'Townsville',
       'BadgerysCreek', 'Tuggeranong', 'Richmond', 'PearceRAAF', 'Launceston',
       'Cobar', 'Moree', 'Mildura', 'Nhil', 'AliceSprings', 'Woomera', 'Uluru',
       'Katherine'],
      dtype='object', name='Location')
```

[67]:
```
len(df1[1].sort_values(ascending = False ).index)
```

[67]:
```
44
```

[71]:
```
location = {'Portland':1, 'Cairns':2, 'NorfolkIsland':3, 'MountGambier':4,␣
 ↪'Walpole':5,
        'Darwin':6, 'NorahHead':7, 'Dartmoor':8, 'CoffsHarbour':9,␣
 ↪'SydneyAirport':10,
        'Ballarat':11, 'Hobart':12, 'GoldCoast':13, 'Witchcliffe':14, 'Watsonia':
 ↪15,
        'Brisbane':16, 'Wollongong':17, 'MelbourneAirport':18, 'Adelaide':19,␣
 ↪'Perth':20,
        'Sydney':21, 'Sale':22, 'PerthAirport':23, 'Nuriootpa':24, 'Albury':25,␣
 ↪'Melbourne':26,
        'Bendigo':27, 'Williamtown':28, 'WaggaWagga':29, 'Canberra':30,␣
 ↪'Townsville':31,
        'BadgerysCreek':32, 'Tuggeranong':33, 'Richmond':34, 'PearceRAAF':35,␣
 ↪'Launceston':36,
        'Cobar':37, 'Moree':38, 'Mildura':39, 'Nhil':40, 'AliceSprings':41,␣
 ↪'Woomera':42, 'Uluru':43,
        'Katherine':44}
df['Location'] = df['Location'].map(location)
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_10968\1194794498.py:10:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df['Location'] = df['Location'].map(location)
```

[73]:
```
df['Date'] = pd.to_datetime(df['Date'], format = '%Y-%m-%dT', errors = 'coerce')
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_10968\3912020841.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df['Date'] = pd.to_datetime(df['Date'], format = '%Y-%m-%dT', errors =
'coerce')
```

[76]:

```
df['Date'].sort_values().
```

[76]:

```
Timestamp('2007-11-01 00:00:00')
```

[ ]:

```
```