In [ ]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
import seaborn as sns
import plotly.express as px
import io
from matplotlib import style
import pathlib
import os
```

In [ ]:

```
data = pd.read_csv("/content/timeseries/timeseries.csv")
```

In [ ]:

```
print('Data First 5 Rows\n')
data.head()
```

```
Data First 5 Rows
```

Out[ ]:

	fips	date	PRECTOT	PS	QV2M	T2M	T2MDEW	T2MWET	T2M_MAX	T2M_MIN	...	TS	WS10M	WS10M_MAX	WS10M_MIN	WS10M_RANGE	WS50M	WS50M_MAX	WS
0	1001	2000-01-01	0.22	100.51	9.65	14.74	13.51	13.51	20.96	11.46	...	14.65	2.20	2.94	1.49	1.46	4.85	6.04	
1	1001	2000-01-02	0.20	100.55	10.42	16.69	14.71	14.71	22.80	12.61	...	16.60	2.52	3.43	1.83	1.60	5.33	6.13	
2	1001	2000-01-03	3.65	100.15	11.76	18.49	16.52	16.52	22.73	15.32	...	18.41	4.03	5.33	2.66	2.67	7.53	9.52	
3	1001	2000-01-04	15.95	100.29	6.42	11.40	6.09	6.10	18.09	2.16	...	11.31	3.84	5.67	2.08	3.59	6.73	9.31	
4	1001	2000-01-05	0.00	101.15	2.95	3.86	-3.29	-3.20	10.82	-2.66	...	2.65	1.60	2.50	0.52	1.98	2.94	4.85	

5 rows × 21 columns

In [ ]:

```
print('Data Last 5 Rows Show\n')
data.tail()
```

```
Data Last 5 Rows Show
```

Out[ ]:

	fips	date	PRECTOT	PS	QV2M	T2M	T2MDEW	T2MWET	T2M_MAX	T2M_MIN	...	TS	WS10M	WS10M_MAX	WS10M_MIN	WS10M_RANGE	WS50M	WS50M_
19300675	56043	2016-12-27	0.16	82.88	1.63	-7.97	-13.49	-12.81	-1.39	-13.60	...	-9.41	5.90	7.63	3.61	4.02	8.58	
19300676	56043	2016-12-28	0.02	83.33	1.41	-8.71	-14.10	-13.84	-2.49	-13.56	...	-10.55	6.50	11.43	4.11	7.32	9.92	
19300677	56043	2016-12-29	0.00	83.75	1.59	-7.96	-13.30	-13.03	0.42	-14.51	...	-10.29	4.29	6.24	2.03	4.22	6.56	

	fips	date	PRECTOT	PS	QV2M	T2M	T2MDEW	T2MWET	T2M_MAX	T2M_MIN	...	TS	WS10M	WS10M_MAX	WS10M_MIN	WS10M_RANGE	WS50M	WS50M_
19300678	56043	2016-12-30	1.22	82.49	2.63	-2.94	-7.40	-7.33	3.76	-6.86	...	-4.14	4.98	7.34	1.99	5.35	7.28	
19300679	56043	2016-12-31	0.44	82.19	1.75	-7.56	-11.98	-11.82	-0.95	-11.61	...	-10.17	2.31	3.47	0.41	3.06	3.37	

5 rows × 21 columns

In [ ]:
```
print('Data Show Describe\n')
data.describe()
```

```
Data Show Describe
```

Out[ ]:

	fips	PRECTOT	PS	QV2M	T2M	T2MDEW	T2MWET	T2M_MAX	T2M_MIN	T2M_RANGE	TS	WS
count	1.930068e+07	1.930068e+07	1.930068e+07	1.930068e+07	1.930068e+07	1.930068e+07	1.930068e+07	1.930068e+07	1.930068e+07	1.930068e+07	1.930068e+07	1.930068
mean	3.067038e+04	2.644145e+00	9.665578e+01	7.816178e+00	1.280146e+01	6.951072e+00	6.986916e+00	1.868141e+01	7.411665e+00	1.126974e+01	1.288900e+01	3.564013
std	1.497911e+04	6.226305e+00	5.447994e+00	4.694305e+00	1.092674e+01	1.014551e+01	1.009116e+01	1.154487e+01	1.057680e+01	4.005165e+00	1.114961e+01	1.862297
min	1.001000e+03	0.000000e+00	6.612000e+01	1.200000e-01	-3.734000e+01	-3.770000e+01	-3.746000e+01	-3.220000e+01	-4.596000e+01	3.000000e-02	-3.823000e+01	2.500000
25%	1.904450e+04	0.000000e+00	9.584000e+01	3.780000e+00	4.450000e+00	-9.600000e-01	-9.200000e-01	1.027000e+01	-7.000000e-01	8.420000e+00	4.340000e+00	2.140000
50%	2.921200e+04	1.800000e-01	9.830000e+01	6.840000e+00	1.402000e+01	7.570000e+00	7.580000e+00	2.040000e+01	8.030000e+00	1.124000e+01	1.404000e+01	3.140000
75%	4.600750e+04	2.160000e+00	9.996000e+01	1.135000e+01	2.188000e+01	1.552000e+01	1.552000e+01	2.787000e+01	1.614000e+01	1.408000e+01	2.207000e+01	4.600000
max	5.604300e+04	2.345900e+02	1.043200e+02	2.292000e+01	4.139000e+01	2.755000e+01	2.755000e+01	4.991000e+01	3.380000e+01	3.461000e+01	4.385000e+01	2.369000

In [ ]:
```
print('Data Show Info\n')
data.info()
```

```
Data Show Info

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19300680 entries, 0 to 19300679
Data columns (total 21 columns):
 #   Column       Dtype  
---  ------       -----  
 0   fips         int64  
 1   date         object 
 2   PRECTOT      float64
 3   PS           float64
 4   QV2M         float64
 5   T2M          float64
 6   T2MDEW       float64
 7   T2MWET       float64
 8   T2M_MAX      float64
 9   T2M_MIN      float64
 10  T2M_RANGE    float64
 11  TS           float64
 12  WS10M        float64
 13  WS10M_MAX    float64
 14  WS10M_MIN    float64
 15  WS10M_RANGE  float64
```

```
 16  WS50M         float64
 17  WS50M_MAX     float64
 18  WS50M_MIN     float64
 19  WS50M_RANGE   float64
 20  score         float64
dtypes: float64(19), int64(1), object(1)
memory usage: 3.0+ GB
```

In [ ]:
```
print('Data Show Columns:\n')
data.columns
```

```
Data Show Columns:
```

Out[ ]:
```
Index(['fips', 'date', 'PRECTOT', 'PS', 'QV2M', 'T2M', 'T2MDEW', 'T2MWET',
       'T2M_MAX', 'T2M_MIN', 'T2M_RANGE', 'TS', 'WS10M', 'WS10M_MAX',
       'WS10M_MIN', 'WS10M_RANGE', 'WS50M', 'WS50M_MAX', 'WS50M_MIN',
       'WS50M_RANGE', 'score'],
      dtype='object')
```

In [ ]:
```
#how many rows and columns are there for all data?
print('Data Shape Show\n')
data.shape
```

```
Data Shape Show
```

Out[ ]:
```
(19300680, 21)
```

In [ ]:
```
print('Data Sum of Null Values \n')
data.isnull().sum()
```

```
Data Sum of Null Values
```

Out[ ]:
```
fips                   0
date                   0
PRECTOT                0
PS                     0
QV2M                   0
T2M                    0
T2MDEW                 0
T2MWET                 0
T2M_MAX                0
T2M_MIN                0
T2M_RANGE              0
TS                     0
WS10M                  0
WS10M_MAX              0
WS10M_MIN              0
WS10M_RANGE            0
WS50M                  0
WS50M_MAX              0
WS50M_MIN              0
WS50M_RANGE            0
score           16543884
dtype: int64
```

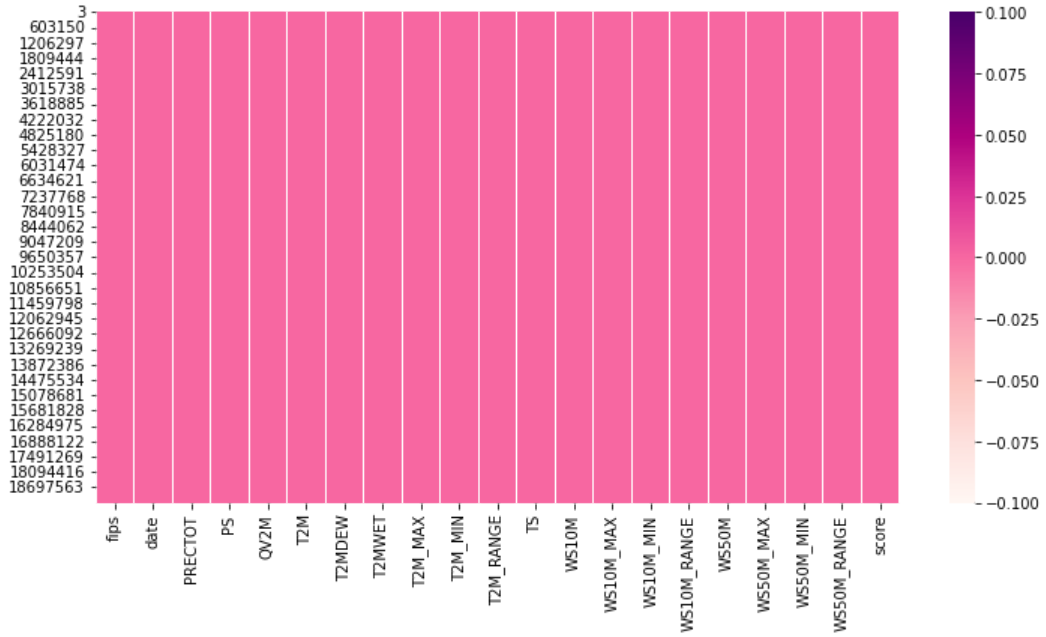
In [ ]:
```
data.isnull().values.any()
```

Out[ ]:
```
True
```

In [ ]:
```
#dropping missing values in target variable
data = data.dropna()
data.isnull().sum()
```

Out[ ]:
```
fips            0
date            0
PRECTOT         0
PS              0
QV2M            0
T2M             0
T2MDEW          0
T2MWET          0
T2M_MAX         0
T2M_MIN         0
T2M_RANGE       0
TS              0
WS10M           0
WS10M_MAX       0
WS10M_MIN       0
WS10M_RANGE     0
WS50M           0
WS50M_MAX       0
WS50M_MIN       0
WS50M_RANGE     0
score           0
dtype: int64
```

In [ ]:
```
plt.figure(figsize=(12,6))
sns.heatmap(data.isnull(),cmap='RdPu')
plt.show()
```



In [ ]:
```
data.dtypes
```

Out[ ]:
```
fips             int64
date            object
PRECTOT        float64
PS             float64
QV2M           float64
T2M            float64
T2MDEW         float64
T2MWET         float64
T2M_MAX        float64
T2M_MIN        float64
T2M_RANGE      float64
TS             float64
WS10M          float64
WS10M_MAX      float64
WS10M_MIN      float64
WS10M_RANGE    float64
WS50M          float64
WS50M_MAX      float64
WS50M_MIN      float64
WS50M_RANGE    float64
score          float64
dtype: object
```

In [ ]:
```
column_list = list(data.columns)
column_list
```

Out[ ]:
```
['fips',
 'date',
 'PRECTOT',
 'PS',
 'QV2M',
 'T2M',
 'T2MDEW',
 'T2MWET',
 'T2M_MAX',
 'T2M_MIN',
 'T2M_RANGE',
 'TS',
 'WS10M',
 'WS10M_MAX',
 'WS10M_MIN',
 'WS10M_RANGE',
 'WS50M',
 'WS50M_MAX',
 'WS50M_MIN',
 'WS50M_RANGE',
 'score']
```

In [ ]:
```
##number of unique values in each of the columns.
data.nunique()
```

Out[ ]:
```
fips            3108
date             887
PRECTOT         6737
PS              3636
QV2M            2165
T2M             6606
T2MDEW          5582
T2MWET          5530
T2M_MAX         6905
T2M_MIN         6557
T2M_RANGE       2787
TS              6928
```

```
WS10M           1513
WS10M_MAX       1877
WS10M_MIN       1276
WS10M_RANGE     1592
WS50M           1854
WS50M_MAX       2286
WS50M_MIN       1572
WS50M_RANGE     1935
score          55395
dtype: int64
```

In [ ]:
```
date = data['date']
date.head()
```

Out[ ]:
```
3     2000-01-04
10    2000-01-11
17    2000-01-18
24    2000-01-25
31    2000-02-01
Name: date, dtype: object
```

In [ ]:
```
#extract year, day and month into new columns
data['year'] = pd.DatetimeIndex(date).year
data['month'] = pd.DatetimeIndex(date).month
data['day'] = pd.DatetimeIndex(date).day
data.dtypes
```

Out[ ]:
```
fips            int64
date           object
PRECTOT       float64
PS            float64
QV2M          float64
T2M           float64
T2MDEW        float64
T2MWET        float64
T2M_MAX       float64
T2M_MIN       float64
T2M_RANGE     float64
TS            float64
WS10M         float64
WS10M_MAX     float64
WS10M_MIN     float64
WS10M_RANGE   float64
WS50M         float64
WS50M_MAX     float64
WS50M_MIN     float64
WS50M_RANGE   float64
score         float64
year            int64
month           int64
day             int64
dtype: object
```

In [ ]:
```
data['score'].value_counts()
```

Out[ ]:
```
0.0000    1480827
1.0000     219135
2.0000     123789
3.0000      82801
4.0000      45841
           ...
```

```
0.1145          1
2.0172          1
0.6750          1
1.3998          1
0.6060          1
Name: score, Length: 55395, dtype: int64
```

In [ ]:
```
#bining target variable into 6classes
data['score'] = data['score'].round().astype(int)
```

In [ ]:
```
data['score'].value_counts()
```

Out[ ]:
```
0    1652230
1     466944
2     295331
3     196802
4     106265
5      39224
Name: score, dtype: int64
```

In [ ]:
```
data.describe()
```

Out[ ]:

	fips	PRECTOT	PS	QV2M	T2M	T2MDEW	T2MWET	T2M_MAX	T2M_MIN	T2M_RANGE	...	WS10M_MIN	WS1C
count	2.756796e+06	2.756796e+06	2.756796e+06	2.756796e+06	2.756796e+06	2.756796e+06	2.756796e+06	2.756796e+06	2.756796e+06	2.756796e+06	...	2.756796e+06	2.7
mean	3.067038e+04	2.714566e+00	9.664736e+01	7.875770e+00	1.289923e+01	7.049350e+00	7.084938e+00	1.876711e+01	7.519916e+00	1.124720e+01	...	1.920655e+00	3.2
std	1.497911e+04	6.247590e+00	5.444698e+00	4.721459e+00	1.097040e+01	1.019765e+01	1.014364e+01	1.160295e+01	1.061818e+01	4.038022e+00	...	1.342458e+00	1.9
min	1.001000e+03	0.000000e+00	6.649000e+01	1.400000e-01	-3.544000e+01	-3.544000e+01	-3.546000e+01	-3.003000e+01	-4.085000e+01	1.600000e-01	...	0.000000e+00	2.6
25%	1.904450e+04	0.000000e+00	9.583000e+01	3.810000e+00	4.580000e+00	-8.800000e-01	-8.400000e-01	1.036000e+01	-5.700000e-01	8.370000e+00	...	9.600000e-01	1.8
50%	2.921200e+04	1.900000e-01	9.828000e+01	6.940000e+00	1.421000e+01	7.810000e+00	7.810000e+00	2.062000e+01	8.260000e+00	1.120000e+01	...	1.660000e+00	2.8
75%	4.600750e+04	2.260000e+00	9.994000e+01	1.145000e+01	2.200000e+01	1.567000e+01	1.567000e+01	2.797000e+01	1.628000e+01	1.408000e+01	...	2.570000e+00	4.2
max	5.604300e+04	1.686900e+02	1.037600e+02	2.212000e+01	3.933000e+01	2.687000e+01	2.687000e+01	4.775000e+01	3.228000e+01	3.017000e+01	...	1.462000e+01	1.8

8 rows × 23 columns

In [ ]:
```
data.describe(include=['object'])
```

Out[ ]:

	date
count	2756796
unique	887
top	2000-01-04
freq	3108

In [ ]:
```
#Removing special characters from continuous features
for c in ['fips', 'date', 'PRECTOT', 'PS', 'QV2M', 'T2M', 'T2MDEW', 'T2MWET',
```

```
        'T2M_MAX', 'T2M_MIN', 'T2M_RANGE', 'TS', 'WS10M', 'WS10M_MAX',
        'WS10M_MIN', 'WS10M_RANGE', 'WS50M', 'WS50M_MAX', 'WS50M_MIN',
        'WS50M_RANGE', 'score']:
    unique_val_cols = data[c].unique()
    print ('Unique values in ' , c , 'are ', unique_val_cols)
```

```
Unique values in  fips are  [ 1001  1003  1005 ... 56039 56041 56043]
Unique values in  date are  ['2000-01-04' '2000-01-11' '2000-01-18' '2000-01-25' '2000-02-01'
 '2000-02-08' '2000-02-15' '2000-02-22' '2000-02-29' '2000-03-07'
 '2000-03-14' '2000-03-21' '2000-03-28' '2000-04-04' '2000-04-11'
 '2000-04-18' '2000-04-25' '2000-05-02' '2000-05-09' '2000-05-16'
 '2000-05-23' '2000-05-30' '2000-06-06' '2000-06-13' '2000-06-20'
 '2000-06-27' '2000-07-04' '2000-07-11' '2000-07-18' '2000-07-25'
 '2000-08-01' '2000-08-08' '2000-08-15' '2000-08-22' '2000-08-29'
 '2000-09-05' '2000-09-12' '2000-09-19' '2000-09-26' '2000-10-03'
 '2000-10-10' '2000-10-17' '2000-10-24' '2000-10-31' '2000-11-07'
 '2000-11-14' '2000-11-21' '2000-11-28' '2000-12-05' '2000-12-12'
 '2000-12-19' '2000-12-26' '2001-01-02' '2001-01-09' '2001-01-16'
 '2001-01-23' '2001-01-30' '2001-02-06' '2001-02-13' '2001-02-20'
 '2001-02-27' '2001-03-06' '2001-03-13' '2001-03-20' '2001-03-27'
 '2001-04-03' '2001-04-10' '2001-04-17' '2001-04-24' '2001-05-01'
 '2001-05-08' '2001-05-15' '2001-05-22' '2001-05-29' '2001-06-05'
 '2001-06-12' '2001-06-19' '2001-06-26' '2001-07-03' '2001-07-10'
 '2001-07-17' '2001-07-24' '2001-07-31' '2001-08-07' '2001-08-14'
 '2001-08-21' '2001-08-28' '2001-09-04' '2001-09-11' '2001-09-18'
 '2001-09-25' '2001-10-02' '2001-10-09' '2001-10-16' '2001-10-23'
 '2001-10-30' '2001-11-06' '2001-11-13' '2001-11-20' '2001-11-27'
 '2001-12-04' '2001-12-11' '2001-12-18' '2001-12-25' '2002-01-01'
 '2002-01-08' '2002-01-15' '2002-01-22' '2002-01-29' '2002-02-05'
 '2002-02-12' '2002-02-19' '2002-02-26' '2002-03-05' '2002-03-12'
 '2002-03-19' '2002-03-26' '2002-04-02' '2002-04-09' '2002-04-16'
 '2002-04-23' '2002-04-30' '2002-05-07' '2002-05-14' '2002-05-21'
 '2002-05-28' '2002-06-04' '2002-06-11' '2002-06-18' '2002-06-25'
 '2002-07-02' '2002-07-09' '2002-07-16' '2002-07-23' '2002-07-30'
 '2002-08-06' '2002-08-13' '2002-08-20' '2002-08-27' '2002-09-03'
 '2002-09-10' '2002-09-17' '2002-09-24' '2002-10-01' '2002-10-08'
 '2002-10-15' '2002-10-22' '2002-10-29' '2002-11-05' '2002-11-12'
 '2002-11-19' '2002-11-26' '2002-12-03' '2002-12-10' '2002-12-17'
 '2002-12-24' '2002-12-31' '2003-01-07' '2003-01-14' '2003-01-21'
 '2003-01-28' '2003-02-04' '2003-02-11' '2003-02-18' '2003-02-25'
 '2003-03-04' '2003-03-11' '2003-03-18' '2003-03-25' '2003-04-01'
 '2003-04-08' '2003-04-15' '2003-04-22' '2003-04-29' '2003-05-06'
 '2003-05-13' '2003-05-20' '2003-05-27' '2003-06-03' '2003-06-10'
 '2003-06-17' '2003-06-24' '2003-07-01' '2003-07-08' '2003-07-15'
 '2003-07-22' '2003-07-29' '2003-08-05' '2003-08-12' '2003-08-19'
 '2003-08-26' '2003-09-02' '2003-09-09' '2003-09-16' '2003-09-23'
 '2003-09-30' '2003-10-07' '2003-10-14' '2003-10-21' '2003-10-28'
 '2003-11-04' '2003-11-11' '2003-11-18' '2003-11-25' '2003-12-02'
 '2003-12-09' '2003-12-16' '2003-12-23' '2003-12-30' '2004-01-06'
 '2004-01-13' '2004-01-20' '2004-01-27' '2004-02-03' '2004-02-10'
 '2004-02-17' '2004-02-24' '2004-03-02' '2004-03-09' '2004-03-16'
 '2004-03-23' '2004-03-30' '2004-04-06' '2004-04-13' '2004-04-20'
 '2004-04-27' '2004-05-04' '2004-05-11' '2004-05-18' '2004-05-25'
 '2004-06-01' '2004-06-08' '2004-06-15' '2004-06-22' '2004-06-29'
 '2004-07-06' '2004-07-13' '2004-07-20' '2004-07-27' '2004-08-03'
 '2004-08-10' '2004-08-17' '2004-08-24' '2004-08-31' '2004-09-07'
 '2004-09-14' '2004-09-21' '2004-09-28' '2004-10-05' '2004-10-12'
 '2004-10-19' '2004-10-26' '2004-11-02' '2004-11-09' '2004-11-16'
 '2004-11-23' '2004-11-30' '2004-12-07' '2004-12-14' '2004-12-21'
 '2004-12-28' '2005-01-04' '2005-01-11' '2005-01-18' '2005-01-25'
 '2005-02-01' '2005-02-08' '2005-02-15' '2005-02-22' '2005-03-01'
 '2005-03-08' '2005-03-15' '2005-03-22' '2005-03-29' '2005-04-05'
 '2005-04-12' '2005-04-19' '2005-04-26' '2005-05-03' '2005-05-10'
 '2005-05-17' '2005-05-24' '2005-05-31' '2005-06-07' '2005-06-14'
 '2005-06-21' '2005-06-28' '2005-07-05' '2005-07-12' '2005-07-19'
```

'2005-07-26' '2005-08-02' '2005-08-09' '2005-08-16' '2005-08-23'
'2005-08-30' '2005-09-06' '2005-09-13' '2005-09-20' '2005-09-27'
'2005-10-04' '2005-10-11' '2005-10-18' '2005-10-25' '2005-11-01'
'2005-11-08' '2005-11-15' '2005-11-22' '2005-11-29' '2005-12-06'
'2005-12-13' '2005-12-20' '2005-12-27' '2006-01-03' '2006-01-10'
'2006-01-17' '2006-01-24' '2006-01-31' '2006-02-07' '2006-02-14'
'2006-02-21' '2006-02-28' '2006-03-07' '2006-03-14' '2006-03-21'
'2006-03-28' '2006-04-04' '2006-04-11' '2006-04-18' '2006-04-25'
'2006-05-02' '2006-05-09' '2006-05-16' '2006-05-23' '2006-05-30'
'2006-06-06' '2006-06-13' '2006-06-20' '2006-06-27' '2006-07-04'
'2006-07-11' '2006-07-18' '2006-07-25' '2006-08-01' '2006-08-08'
'2006-08-15' '2006-08-22' '2006-08-29' '2006-09-05' '2006-09-12'
'2006-09-19' '2006-09-26' '2006-10-03' '2006-10-10' '2006-10-17'
'2006-10-24' '2006-10-31' '2006-11-07' '2006-11-14' '2006-11-21'
'2006-11-28' '2006-12-05' '2006-12-12' '2006-12-19' '2006-12-26'
'2007-01-02' '2007-01-09' '2007-01-16' '2007-01-23' '2007-01-30'
'2007-02-06' '2007-02-13' '2007-02-20' '2007-02-27' '2007-03-06'
'2007-03-13' '2007-03-20' '2007-03-27' '2007-04-03' '2007-04-10'
'2007-04-17' '2007-04-24' '2007-05-01' '2007-05-08' '2007-05-15'
'2007-05-22' '2007-05-29' '2007-06-05' '2007-06-12' '2007-06-19'
'2007-06-26' '2007-07-03' '2007-07-10' '2007-07-17' '2007-07-24'
'2007-07-31' '2007-08-07' '2007-08-14' '2007-08-21' '2007-08-28'
'2007-09-04' '2007-09-11' '2007-09-18' '2007-09-25' '2007-10-02'
'2007-10-09' '2007-10-16' '2007-10-23' '2007-10-30' '2007-11-06'
'2007-11-13' '2007-11-20' '2007-11-27' '2007-12-04' '2007-12-11'
'2007-12-18' '2007-12-25' '2008-01-01' '2008-01-08' '2008-01-15'
'2008-01-22' '2008-01-29' '2008-02-05' '2008-02-12' '2008-02-19'
'2008-02-26' '2008-03-04' '2008-03-11' '2008-03-18' '2008-03-25'
'2008-04-01' '2008-04-08' '2008-04-15' '2008-04-22' '2008-04-29'
'2008-05-06' '2008-05-13' '2008-05-20' '2008-05-27' '2008-06-03'
'2008-06-10' '2008-06-17' '2008-06-24' '2008-07-01' '2008-07-08'
'2008-07-15' '2008-07-22' '2008-07-29' '2008-08-05' '2008-08-12'
'2008-08-19' '2008-08-26' '2008-09-02' '2008-09-09' '2008-09-16'
'2008-09-23' '2008-09-30' '2008-10-07' '2008-10-14' '2008-10-21'
'2008-10-28' '2008-11-04' '2008-11-11' '2008-11-18' '2008-11-25'
'2008-12-02' '2008-12-09' '2008-12-16' '2008-12-23' '2008-12-30'
'2009-01-06' '2009-01-13' '2009-01-20' '2009-01-27' '2009-02-03'
'2009-02-10' '2009-02-17' '2009-02-24' '2009-03-03' '2009-03-10'
'2009-03-17' '2009-03-24' '2009-03-31' '2009-04-07' '2009-04-14'
'2009-04-21' '2009-04-28' '2009-05-05' '2009-05-12' '2009-05-19'
'2009-05-26' '2009-06-02' '2009-06-09' '2009-06-16' '2009-06-23'
'2009-06-30' '2009-07-07' '2009-07-14' '2009-07-21' '2009-07-28'
'2009-08-04' '2009-08-11' '2009-08-18' '2009-08-25' '2009-09-01'
'2009-09-08' '2009-09-15' '2009-09-22' '2009-09-29' '2009-10-06'
'2009-10-13' '2009-10-20' '2009-10-27' '2009-11-03' '2009-11-10'
'2009-11-17' '2009-11-24' '2009-12-01' '2009-12-08' '2009-12-15'
'2009-12-22' '2009-12-29' '2010-01-05' '2010-01-12' '2010-01-19'
'2010-01-26' '2010-02-02' '2010-02-09' '2010-02-16' '2010-02-23'
'2010-03-02' '2010-03-09' '2010-03-16' '2010-03-23' '2010-03-30'
'2010-04-06' '2010-04-13' '2010-04-20' '2010-04-27' '2010-05-04'
'2010-05-11' '2010-05-18' '2010-05-25' '2010-06-01' '2010-06-08'
'2010-06-15' '2010-06-22' '2010-06-29' '2010-07-06' '2010-07-13'
'2010-07-20' '2010-07-27' '2010-08-03' '2010-08-10' '2010-08-17'
'2010-08-24' '2010-08-31' '2010-09-07' '2010-09-14' '2010-09-21'
'2010-09-28' '2010-10-05' '2010-10-12' '2010-10-19' '2010-10-26'
'2010-11-02' '2010-11-09' '2010-11-16' '2010-11-23' '2010-11-30'
'2010-12-07' '2010-12-14' '2010-12-21' '2010-12-28' '2011-01-04'
'2011-01-11' '2011-01-18' '2011-01-25' '2011-02-01' '2011-02-08'
'2011-02-15' '2011-02-22' '2011-03-01' '2011-03-08' '2011-03-15'
'2011-03-22' '2011-03-29' '2011-04-05' '2011-04-12' '2011-04-19'
'2011-04-26' '2011-05-03' '2011-05-10' '2011-05-17' '2011-05-24'
'2011-05-31' '2011-06-07' '2011-06-14' '2011-06-21' '2011-06-28'
'2011-07-05' '2011-07-12' '2011-07-19' '2011-07-26' '2011-08-02'
'2011-08-09' '2011-08-16' '2011-08-23' '2011-08-30' '2011-09-06'
'2011-09-13' '2011-09-20' '2011-09-27' '2011-10-04' '2011-10-11'

```
 '2011-10-18' '2011-10-25' '2011-11-01' '2011-11-08' '2011-11-15'
 '2011-11-22' '2011-11-29' '2011-12-06' '2011-12-13' '2011-12-20'
 '2011-12-27' '2012-01-03' '2012-01-10' '2012-01-17' '2012-01-24'
 '2012-01-31' '2012-02-07' '2012-02-14' '2012-02-21' '2012-02-28'
 '2012-03-06' '2012-03-13' '2012-03-20' '2012-03-27' '2012-04-03'
 '2012-04-10' '2012-04-17' '2012-04-24' '2012-05-01' '2012-05-08'
 '2012-05-15' '2012-05-22' '2012-05-29' '2012-06-05' '2012-06-12'
 '2012-06-19' '2012-06-26' '2012-07-03' '2012-07-10' '2012-07-17'
 '2012-07-24' '2012-07-31' '2012-08-07' '2012-08-14' '2012-08-21'
 '2012-08-28' '2012-09-04' '2012-09-11' '2012-09-18' '2012-09-25'
 '2012-10-02' '2012-10-09' '2012-10-16' '2012-10-23' '2012-10-30'
 '2012-11-06' '2012-11-13' '2012-11-20' '2012-11-27' '2012-12-04'
 '2012-12-11' '2012-12-18' '2012-12-25' '2013-01-01' '2013-01-08'
 '2013-01-15' '2013-01-22' '2013-01-29' '2013-02-05' '2013-02-12'
 '2013-02-19' '2013-02-26' '2013-03-05' '2013-03-12' '2013-03-19'
 '2013-03-26' '2013-04-02' '2013-04-09' '2013-04-16' '2013-04-23'
 '2013-04-30' '2013-05-07' '2013-05-14' '2013-05-21' '2013-05-28'
 '2013-06-04' '2013-06-11' '2013-06-18' '2013-06-25' '2013-07-02'
 '2013-07-09' '2013-07-16' '2013-07-23' '2013-07-30' '2013-08-06'
 '2013-08-13' '2013-08-20' '2013-08-27' '2013-09-03' '2013-09-10'
 '2013-09-17' '2013-09-24' '2013-10-01' '2013-10-08' '2013-10-15'
 '2013-10-22' '2013-10-29' '2013-11-05' '2013-11-12' '2013-11-19'
 '2013-11-26' '2013-12-03' '2013-12-10' '2013-12-17' '2013-12-24'
 '2013-12-31' '2014-01-07' '2014-01-14' '2014-01-21' '2014-01-28'
 '2014-02-04' '2014-02-11' '2014-02-18' '2014-02-25' '2014-03-04'
 '2014-03-11' '2014-03-18' '2014-03-25' '2014-04-01' '2014-04-08'
 '2014-04-15' '2014-04-22' '2014-04-29' '2014-05-06' '2014-05-13'
 '2014-05-20' '2014-05-27' '2014-06-03' '2014-06-10' '2014-06-17'
 '2014-06-24' '2014-07-01' '2014-07-08' '2014-07-15' '2014-07-22'
 '2014-07-29' '2014-08-05' '2014-08-12' '2014-08-19' '2014-08-26'
 '2014-09-02' '2014-09-09' '2014-09-16' '2014-09-23' '2014-09-30'
 '2014-10-07' '2014-10-14' '2014-10-21' '2014-10-28' '2014-11-04'
 '2014-11-11' '2014-11-18' '2014-11-25' '2014-12-02' '2014-12-09'
 '2014-12-16' '2014-12-23' '2014-12-30' '2015-01-06' '2015-01-13'
 '2015-01-20' '2015-01-27' '2015-02-03' '2015-02-10' '2015-02-17'
 '2015-02-24' '2015-03-03' '2015-03-10' '2015-03-17' '2015-03-24'
 '2015-03-31' '2015-04-07' '2015-04-14' '2015-04-21' '2015-04-28'
 '2015-05-05' '2015-05-12' '2015-05-19' '2015-05-26' '2015-06-02'
 '2015-06-09' '2015-06-16' '2015-06-23' '2015-06-30' '2015-07-07'
 '2015-07-14' '2015-07-21' '2015-07-28' '2015-08-04' '2015-08-11'
 '2015-08-18' '2015-08-25' '2015-09-01' '2015-09-08' '2015-09-15'
 '2015-09-22' '2015-09-29' '2015-10-06' '2015-10-13' '2015-10-20'
 '2015-10-27' '2015-11-03' '2015-11-10' '2015-11-17' '2015-11-24'
 '2015-12-01' '2015-12-08' '2015-12-15' '2015-12-22' '2015-12-29'
 '2016-01-05' '2016-01-12' '2016-01-19' '2016-01-26' '2016-02-02'
 '2016-02-09' '2016-02-16' '2016-02-23' '2016-03-01' '2016-03-08'
 '2016-03-15' '2016-03-22' '2016-03-29' '2016-04-05' '2016-04-12'
 '2016-04-19' '2016-04-26' '2016-05-03' '2016-05-10' '2016-05-17'
 '2016-05-24' '2016-05-31' '2016-06-07' '2016-06-14' '2016-06-21'
 '2016-06-28' '2016-07-05' '2016-07-12' '2016-07-19' '2016-07-26'
 '2016-08-02' '2016-08-09' '2016-08-16' '2016-08-23' '2016-08-30'
 '2016-09-06' '2016-09-13' '2016-09-20' '2016-09-27' '2016-10-04'
 '2016-10-11' '2016-10-18' '2016-10-25' '2016-11-01' '2016-11-08'
 '2016-11-15' '2016-11-22' '2016-11-29' '2016-12-06' '2016-12-13'
 '2016-12-20' '2016-12-27']
Unique values in  PRECTOT are  [15.95  1.33  1.11 ... 48.91 79.15 59.08]
Unique values in  PS are  [100.29 100.4  100.39 ... 103.65 103.75 103.76]
Unique values in  QV2M are  [ 6.42  6.63  9.53 ... 21.54 21.64 21.44]
Unique values in  T2M are  [ 11.4   11.48  14.28 ... -29.26 -28.93 -26.79]
Unique values in  T2MDEW are  [  6.09   7.84  13.26 ... -29.93 -30.28 -29.54]
Unique values in  T2MWET are  [  6.1    7.84  13.26 ... -29.91 -29.41 -27.54]
Unique values in  T2M_MAX are  [ 18.09  18.88  18.04 ... -20.12 -24.42 -19.53]
Unique values in  T2M_MIN are  [  2.16   5.72   8.98 ... -34.78 -35.15 -35.38]
Unique values in  T2M_RANGE are  [15.92 13.16  9.06 ... 25.4  28.57 28.55]
Unique values in  TS are  [ 11.31  10.43  14.19 ... -26.42 -30.38 -28.05]
```

```
Unique values in  WS10M are  [ 3.84  1.76  2.63 ... 13.75 16.06 14.72]
Unique values in  WS10M_MAX are  [ 5.67  2.48  3.6  ... 19.18 19.94 18.13]
Unique values in  WS10M_MIN are  [ 2.08  1.05  1.67 ... 11.64 12.41 12.53]
Unique values in  WS10M_RANGE are  [ 3.59  1.43  1.92 ... 17.21 17.09 16.89]
Unique values in  WS50M are  [ 6.73  3.55  5.19 ... 19.05 18.39 17.95]
Unique values in  WS50M_MAX are  [ 9.31  6.38  6.4  ...  1.1   1.17 21.98]
Unique values in  WS50M_MIN are  [ 3.74  1.71  3.84 ... 16.58 14.83 15.86]
Unique values in  WS50M_RANGE are  [ 5.58  4.67  2.55 ... 20.41 19.34 19.85]
Unique values in  score are  [1 2 3 4 5 0]
```

In [ ]:

```
for c in ['fips', 'date', 'PRECTOT', 'PS', 'QV2M', 'T2M', 'T2MDEW', 'T2MWET',
       'T2M_MAX', 'T2M_MIN', 'T2M_RANGE', 'TS', 'WS10M', 'WS10M_MAX',
       'WS10M_MIN', 'WS10M_RANGE', 'WS50M', 'WS50M_MAX', 'WS50M_MIN',
       'WS50M_RANGE', 'score']:
  data[c] = pd.to_numeric(data[c], errors='coerce')
```

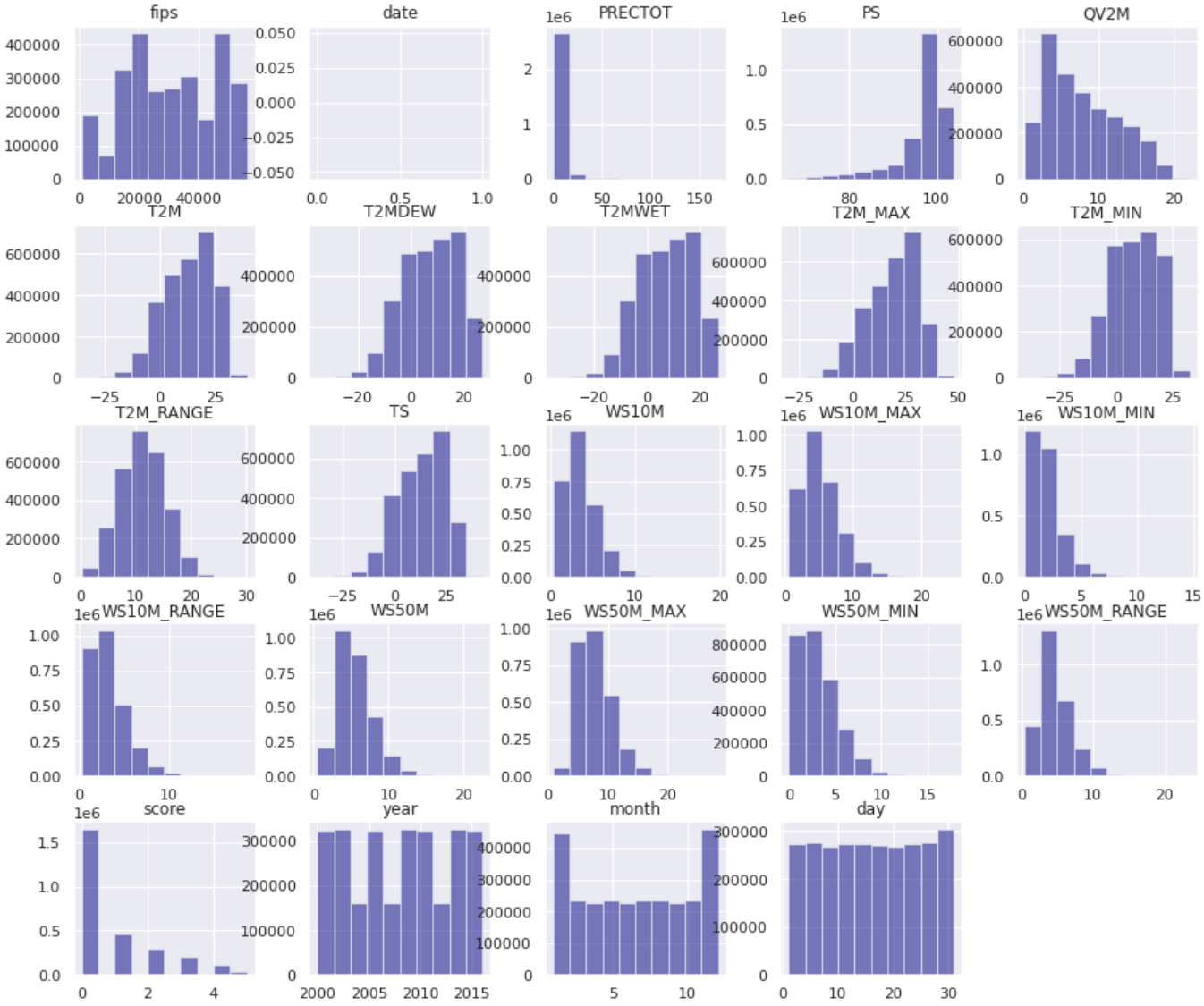
In [ ]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2756796 entries, 3 to 19300675
Data columns (total 24 columns):
 #   Column       Dtype
---  ------       -----
 0   fips         int64
 1   date         float64
 2   PRECTOT      float64
 3   PS           float64
 4   QV2M         float64
 5   T2M          float64
 6   T2MDEW       float64
 7   T2MWET       float64
 8   T2M_MAX      float64
 9   T2M_MIN      float64
 10  T2M_RANGE    float64
 11  TS           float64
 12  WS10M        float64
 13  WS10M_MAX    float64
 14  WS10M_MIN    float64
 15  WS10M_RANGE  float64
 16  WS50M        float64
 17  WS50M_MAX    float64
 18  WS50M_MIN    float64
 19  WS50M_RANGE  float64
 20  score        int64
 21  year         int64
 22  month        int64
 23  day          int64
dtypes: float64(19), int64(5)
memory usage: 525.8 MB
```

In [ ]:

```
sns.set(style="darkgrid")
data.hist(bins=10,figsize=(15,13) ,color = 'navy', alpha = 0.5)
plt.show()
```



The Drought dataset is a labelled dataset. Distribution of scores is analyzed to identify if data is biased or not.

It can be seen that the features PRICTOT ,WS10M-MIN,WS50M-MIN,WS10M-RANGE are skewed to the left.

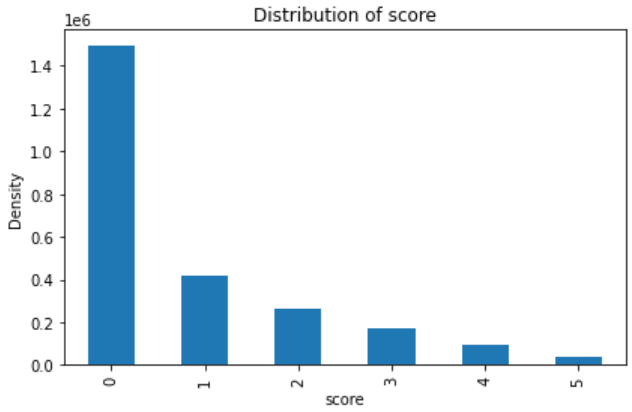
The features PS,T2M,T2M-MAX are skewed to the right while remaining features are fairly well distributed across all range.

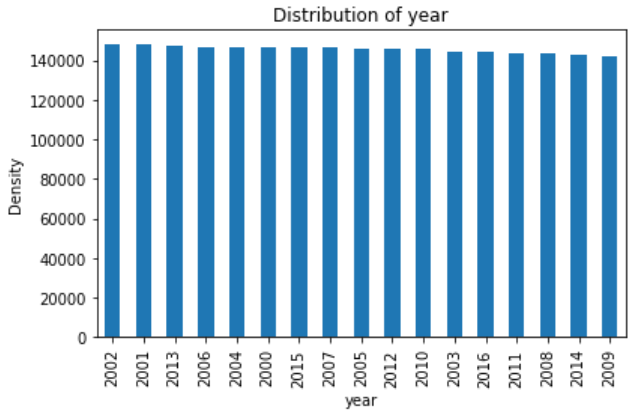
Univariate analysis on categorical variables

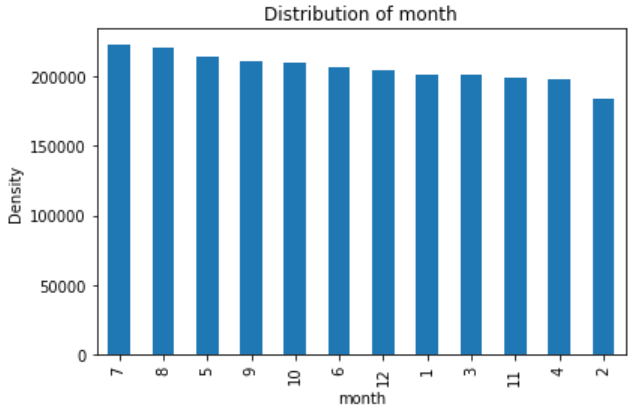
In [ ]:
```
categorical_columns = ['score','year','month','day']
categorical_data = data[['score','year','month','day']]
```

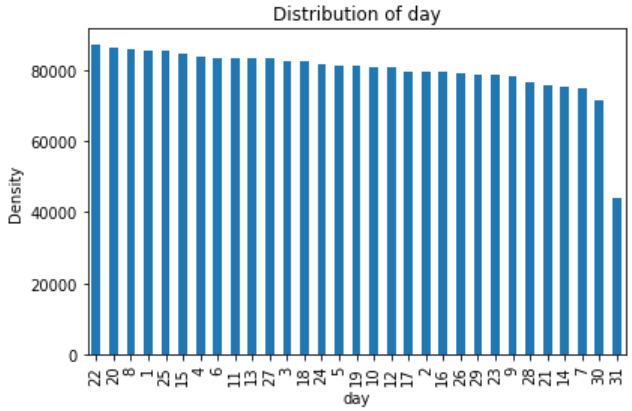
In [ ]:
```
#categorical variables
plt.figure(figsize=(10,40))
for col in categorical_columns:
    plt.figure()
    categorical_data[col].value_counts().plot(kind = 'bar')
    x_label = col
    y_label = 'density'
    plt.ylabel(y_label)
    plt.xlabel(x_label)
    plt.title('Distribution of {x_name}'.format(x_name=x_label))
    plt.tight_layout()
```

```
<Figure size 720x2880 with 0 Axes>
```



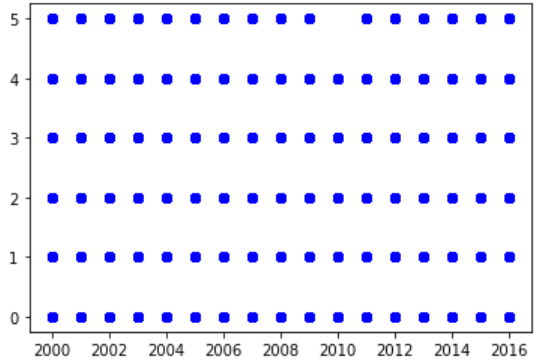






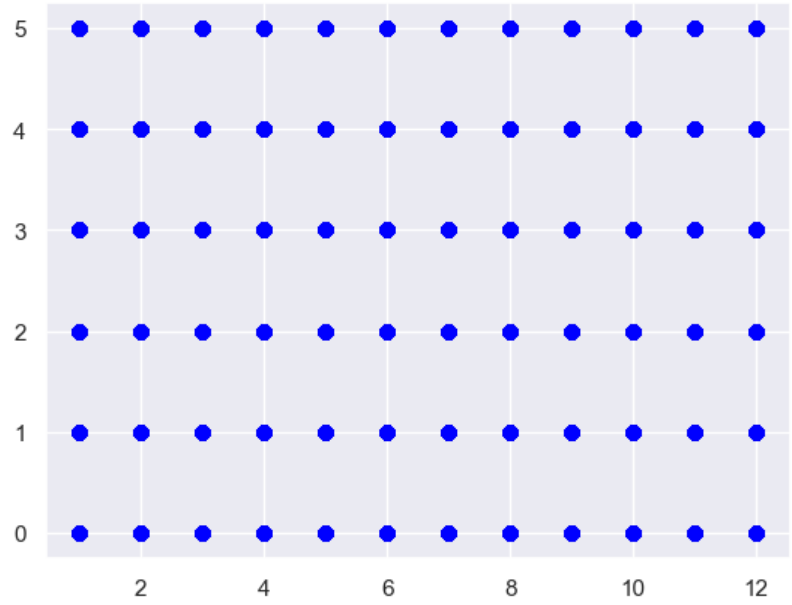
Bivariate analysis

In [ ]:
```
plt.scatter(data['year'], data['score'], c ="blue")
plt.show()
```



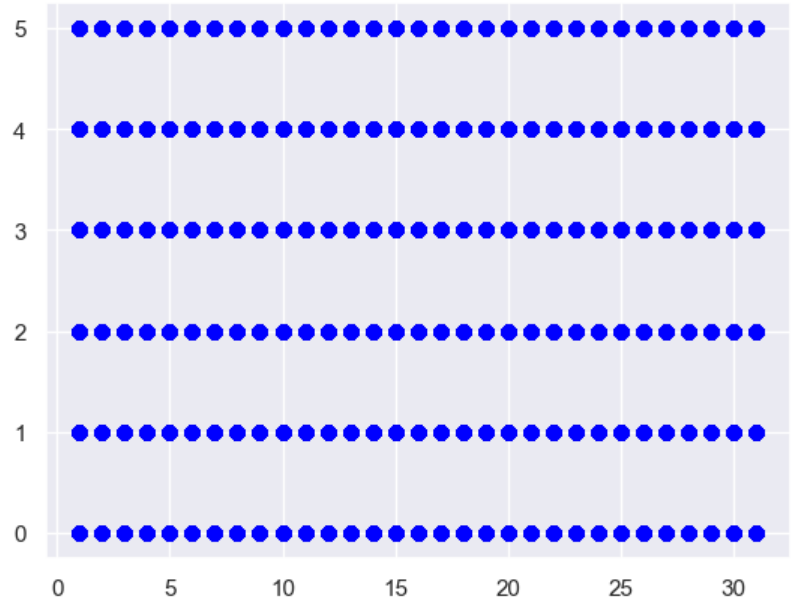
In [ ]:
```
plt.scatter(data['month'], data['score'], c ="blue")
```

```
plt.show()
```



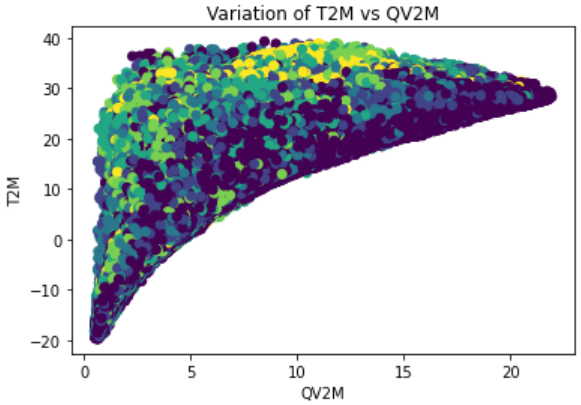
In [ ]:

```
plt.scatter(data['day'], data['score'], c ="blue")
plt.show()
```



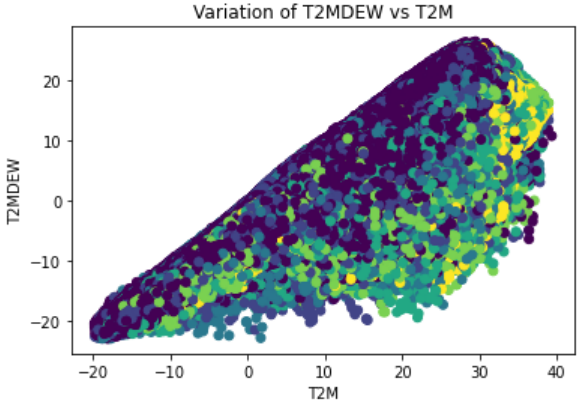
In [ ]:

```
plt.scatter(data['QV2M'], data['T2M'], c =data['score'])
plt.xlabel('QV2M')
plt.ylabel('T2M')
plt.title('QV2M vs T2M')
plt.show()
```



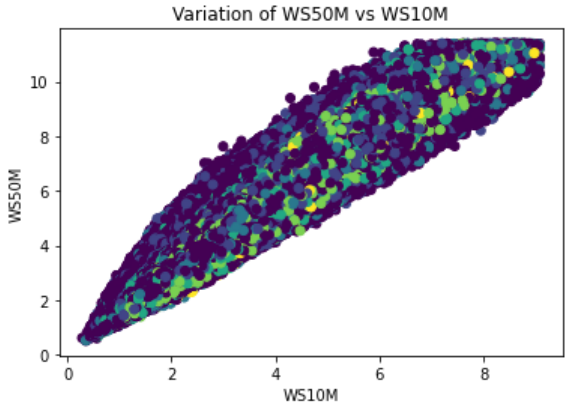
In [ ]:

```
plt.scatter(data['T2M'], data['T2MDEW'], c =data['score'])
plt.xlabel('T2M')
plt.ylabel('T2MDEW')
plt.title('T2M vs T2MDEW')
plt.show()
```



In [ ]:

```
temp_df = data[data['score']==5]
plt.scatter(data['WS10M'], data['WS50M'], c= data['score'])
plt.xlabel('WS10M')
plt.ylabel('WS50M')
plt.title('WS10M vs WS50M')
plt.show()
```



To understand the correlations between features scatter plots were drawn to find out the attributes having strong correlation.

In the above scatter plots we can observe that the independent variables have shown strong positive correlation.

The pairs WS10M - WS50M have the one-to-one relationship.

The features T2M – T2MDEW and QV2M - T2M relationship is not inear, although we can say that the overall they have strong correlation.

Stripping target values from dataset

In [ ]:
```
independent_variables = data.drop('score', 1)
independent_variables = independent_variables.drop('fips', 1)
independent_variables = independent_variables.drop('date', 1)
independent_variables.head()
```

Out[ ]:

	PRECTOT	PS	QV2M	T2M	T2MDEW	T2MWET	T2M_MAX	T2M_MIN	T2M_RANGE	TS	...	WS10M_MAX	WS10M_MIN	WS10M_RANGE	WS50M	WS50M_MAX	WS50M_MI
3	15.95	100.29	6.42	11.40	6.09	6.10	18.09	2.16	15.92	11.31	...	5.67	2.08	3.59	6.73	9.31	3.7
10	1.33	100.40	6.63	11.48	7.84	7.84	18.88	5.72	13.16	10.43	...	2.48	1.05	1.43	3.55	6.38	1.7
17	1.11	100.39	9.53	14.28	13.26	13.26	18.04	8.98	9.06	14.19	...	3.60	1.67	1.92	5.19	6.40	3.8
24	0.00	100.11	2.05	-0.78	-7.93	-7.72	5.65	-5.46	11.11	-0.61	...	4.59	2.28	2.32	5.75	8.03	3.9
31	0.00	101.00	3.36	2.06	-1.73	-1.70	11.02	-4.21	15.23	1.88	...	2.74	0.88	1.86	4.18	6.38	1.2

5 rows × 21 columns

In [ ]:
```
target = data['score']
target.head()
```

Out[ ]:
```
3     1
10    2
17    2
24    2
31    1
Name: score, dtype: int64
```

Correlation between features

In [ ]:
```
list_of_columns = ['PRECTOT','PS','QV2M','T2M','T2MDEW','T2MWET','T2M_MAX','T2M_MIN','T2M_RANGE','TS',
                   'WS10M','WS10M_MAX','WS10M_MIN','WS10M_RANGE','WS50M','WS50M_MAX','WS50M_MIN','WS50M_RANGE']
drought_data_columns = data[['PRECTOT','PS','QV2M','T2M','T2MDEW','T2MWET','T2M_MAX','T2M_MIN',
                             'T2M_RANGE','TS','WS10M','WS10M_MAX','WS10M_MIN','WS10M_RANGE','WS50M'
                            ,'WS50M_MAX','WS50M_MIN','WS50M_RANGE']]
```

In [ ]:
```
correlation_plot = drought_data_columns.corr()
correlation_plot.style.background_gradient(cmap = 'YlGnBu')
```

Out[ ]:

	PRECTOT	PS	QV2M	T2M	T2MDEW	T2MWET	T2M_MAX	T2M_MIN	T2M_RANGE	TS	WS10M	WS10M_MAX	WS10M_MIN	WS10M_RANG
PRECTOT	1.000000	0.068775	0.245081	0.093258	0.231035	0.230975	0.026773	0.144929	-0.304171	0.089598	0.049730	0.060981	0.023346	0.0657
PS	0.068775	1.000000	0.282412	0.164160	0.341234	0.341252	0.111979	0.208285	-0.225935	0.163830	-0.080747	-0.135905	0.022932	-0.1983
QV2M	0.245081	0.282412	1.000000	0.870242	0.959385	0.960434	0.804338	0.906144	-0.071547	0.862559	-0.225449	-0.256452	-0.108789	-0.2692
T2M	0.093258	0.164160	0.870242	1.000000	0.913530	0.914218	0.983356	0.981629	0.244357	0.997515	-0.207874	-0.220192	-0.125407	-0.2090
T2MDEW	0.231035	0.341234	0.959385	0.913530	1.000000	0.999970	0.854716	0.939934	-0.015643	0.905184	-0.238299	-0.268686	-0.115920	-0.2807
T2MWET	0.230975	0.341252	0.960434	0.914218	0.999970	1.000000	0.855401	0.940629	-0.015500	0.905911	-0.237971	-0.268292	-0.115882	-0.2801
T2M_MAX	0.026773	0.111979	0.804338	0.983356	0.854716	0.855401	1.000000	0.937762	0.407534	0.980101	-0.216764	-0.221671	-0.141911	-0.1996
T2M_MIN	0.144929	0.208285	0.906144	0.981629	0.939934	0.940629	0.937762	1.000000	0.065037	0.979134	-0.206382	-0.225829	-0.112878	-0.2252
T2M_RANGE	-0.304171	-0.225935	-0.071547	0.244357	-0.015643	-0.015500	0.407534	0.065037	1.000000	0.241564	-0.080163	-0.043127	-0.110952	0.0187
TS	0.089598	0.163830	0.862559	0.997515	0.905184	0.905911	0.980101	0.979134	0.241564	1.000000	-0.189823	-0.202713	-0.110273	-0.1960
WS10M	0.049730	-0.080747	-0.225449	-0.207874	-0.238299	-0.237971	-0.216764	-0.206382	-0.080163	-0.189823	1.000000	0.952217	0.833340	0.7028
WS10M_MAX	0.060981	-0.135905	-0.256452	-0.220192	-0.268686	-0.268292	-0.221671	-0.225829	-0.043127	-0.202713	0.952217	1.000000	0.690087	0.8660
WS10M_MIN	0.023346	0.022932	-0.108789	-0.125407	-0.115920	-0.115882	-0.141911	-0.112878	-0.110952	-0.110273	0.833340	0.690087	1.000000	0.2357
WS10M_RANGE	0.065755	-0.198332	-0.269203	-0.209030	-0.280702	-0.280199	-0.199614	-0.225256	0.018746	-0.196015	0.702896	0.866026	0.235775	1.0000
WS50M	0.069057	-0.043315	-0.205971	-0.193196	-0.204238	-0.204143	-0.195727	-0.197991	-0.041778	-0.180665	0.966275	0.910717	0.839187	0.6431
WS50M_MAX	0.079508	-0.091821	-0.249961	-0.206444	-0.245323	-0.245147	-0.196236	-0.225744	0.029737	-0.193347	0.908750	0.946710	0.666629	0.8106
WS50M_MIN	0.057816	0.036238	-0.081554	-0.112579	-0.082416	-0.082497	-0.133234	-0.096593	-0.128844	-0.102367	0.795424	0.660428	0.943983	0.2346
WS50M_RANGE	0.047477	-0.154479	-0.246203	-0.159589	-0.239335	-0.239029	-0.126331	-0.200157	0.163320	-0.152434	0.412412	0.592380	-0.046209	0.8273

Attributes T2M_MAX, T2M_MIN, T2MDEW, T2MWET, QV2M, T2M, and TS have shown strong positive correlation

Likewise, WS10M, WS10M_MAX and WS10M_MIN have shown a strong positve correlation

Similarly, WS50M, WS50M_MAX and WS50M_MIN show strong positive correlation

But, from the scatter plots above, we see significant variance between the data points, despite the strong positive correlation. Hence,all the variables are retained, and other feature selection methods needs to be experimented.

In [ ]: