Rudy Fasano

Applied Data Mining

Assignment 1.2 (EDA)

In [1]:
```
import pandas as pd
import matplotlib as mpl
import matplotlib.pyplot as plt
import scipy.stats
import seaborn as sns; sns.set_theme(color_codes=True)
from datetime import date
from datetime import datetime
from numpy import median
%matplotlib inline
```

In [2]:
```
df = pd.read_csv("D:/School/502/Week1/datasets/Website Data Sets/Precipitation Data Dataset.csv")
```

Basic Data Exploration

In [3]:
```
df.head(5)
```

Out[3]:

	DATE	PRCP
0	1/1/2001	0.00
1	1/2/2001	0.00
2	1/3/2001	0.00
3	1/4/2001	0.04
4	1/5/2001	0.14

In [4]:
```
df.shape ## identifying dimensions of the dataset
```

Out[4]:
```
(6191, 2)
```

In [5]:
```
df.info() ## identifying data types and other relevant information
          ## both object and float datatypes were identified
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6191 entries, 0 to 6190
```

```
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   DATE    6191 non-null   object
 1   PRCP    6191 non-null   float64
dtypes: float64(1), object(1)
memory usage: 96.9+ KB
```

In [6]:
```
df.describe() ## establish a screenshot of the data and its simple
statistics
```

Out[6]:

	PRCP
count	6191.000000
mean	0.081197
std	0.230326
min	0.000000
25%	0.000000
50%	0.000000
75%	0.030000
max	4.380000

Exploring data to discover missing values, duplicate values and outliers

In [7]:
```
df.isnull().sum() ### searching for missing values
                  ### none were identified
```

Out[7]:
```
DATE    0
PRCP    0
dtype: int64
```

In [8]:
```
duplicate = df.duplicated()  ### searching for duplicate values
print(duplicate.sum())       ### none were found
df[duplicate]
```

```
0
```

Out[8]:

	DATE	PRCP

Convert object datatype from 'DATE' column to datetime objects in order to plot and find any relative relationship

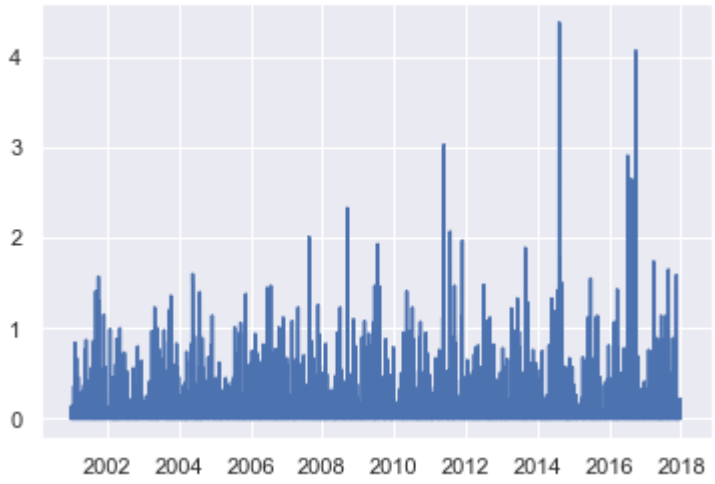
In [12]:
```
# convert dates to datetime to investigate via visualizations
x_values = df['DATE'] = pd.to_datetime(df['DATE'])
y_values = df['PRCP']

df.info() ## verify conversion to datetime object
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6191 entries, 0 to 6190
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype         
---  ------  --------------  -----         
 0   DATE    6191 non-null   datetime64[ns]
 1   PRCP    6191 non-null   float64       
dtypes: datetime64[ns](1), float64(1)
memory usage: 96.9 KB
```

In [17]:
```
plt.plot(x_values, y_values)
```

Out[17]: [<matplotlib.lines.Line2D at 0x1abc2b8a070>]



We can deduce after conducting EDA that the dataset did not have any inconsistencies or human error. We also can deduce that precipitation has steadily increased throughout the first two decades of the new millennium.