```
# This is formatted as code
```

descriptive COVID statistics analysis - New York State

Yannan Li

2/19/2021

Data Import

```
import pandas as pd
import numpy as np
from functools import reduce
import matplotlib.pyplot as plt
%matplotlib inline
```

Loading csv data to notebook

```
covid = pd.read_csv('/content/New_York_State_Statewide_COVID-19_Testing.csv', header=0)
covid.head(5)
```

	Test Date	County	New Positives	Cumulative Number of Positives	Total Number of Tests Performed	Cumulative Number of Tests Performed
0	03/01/2020	Albany	0	0	0	0
1	03/02/2020	Albany	0	0	0	0
2	03/03/2020	Albany	0	0	0	0
3	03/04/2020	Albany	0	0	0	0

Data Cleansing

```
covid1 = covid[covid['County']=="New York"]
covid1.head()
covid1.tail(1)
```

	Test Date	County	New Positives	Cumulative Number of Positives	Total Number of Tests Performed	Cumulative Number of Tests Performed

Data analysis and visualiztinon

1. test for data consistency

```
covid1.sum(axis=0)
```

```
Test Date                             03/01/202003/02/202003/03/202003/04/202003/05/...
County                                New YorkNew YorkNew YorkNew YorkNew YorkNew Yo...
New Positives                                                                     93635
Cumulative Number of Positives                                                 12183438
Total Number of Tests Performed                                                 3496585
Cumulative Number of Tests Performed                                          364588426
dtype: object
```

2. Find dates with top 10 number of cases

```
covid_top = covid1.nlargest(10,"New Positives")
covid_top
```

	Test Date	County	New Positives	Cumulative Number of Positives	Total Number of Tests Performed	Cumulative Number of Tests Performed
10424	04/14/2020	New York	1737	16617	4068	40517
10699	01/14/2021	New York	1066	74215	32603	2927703
10700	01/15/2021	New York	997	75212	29886	2957589
10692	01/07/2021	New York	978	68254	26143	2760768
10706	01/21/2021	New York	932	79812	28980	3082723
10712	01/27/2021	New York	923	84857	26040	3229328
		New				

We can tell from the data that covid cases has two periods of peaks, one is around April 2020, and the second one is around January 2021.

3. Create positive rate and do descriptive analysis

```
# generate positive rate column
covid1["Positive Rate"] = covid1['New Positives'] / covid1['Total Number of Tests Performed']*100

# descriptive analysis
covid1.mean()
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.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-
```

```
New Positives                            2.706214e+02
Cumulative Number of Positives           3.521225e+04
Total Number of Tests Performed          1.010574e+04
Cumulative Number of Tests Performed     1.053724e+06
Positive Rate                            6.602876e+00
dtype: float64
```

```
covid1.std()
```

```
New Positives                            2.781760e+02
Cumulative Number of Positives           2.066696e+04
Total Number of Tests Performed          7.809706e+03
Cumulative Number of Tests Performed     1.025322e+06
Positive Rate                            1.201236e+01
dtype: float64
```

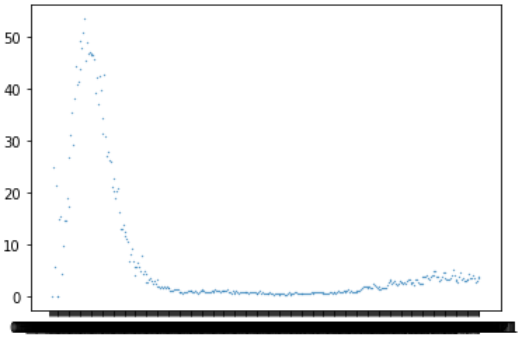
```
covid1.median()
```

```
New Positives                              117.500000
Cumulative Number of Positives           31684.500000
Total Number of Tests Performed           7862.500000
Cumulative Number of Tests Performed    727016.500000
Positive Rate                                1.715802
dtype: float64
```

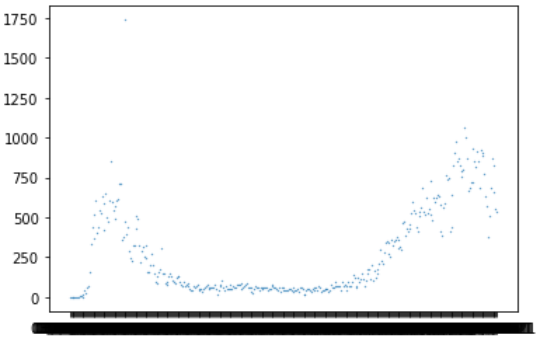
4. Plot the data

```
# plot positive rate over time
date_list = covid1['Test Date'].tolist()
positiver_list = covid1['Positive Rate'].tolist()
```

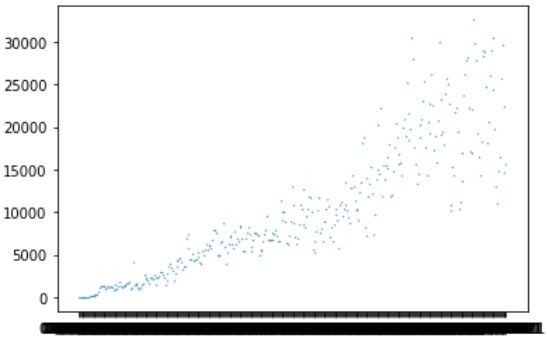
```
# plot positive rate over time
p1 = plt.scatter(date_list,positiver_list, s=0.1);
```



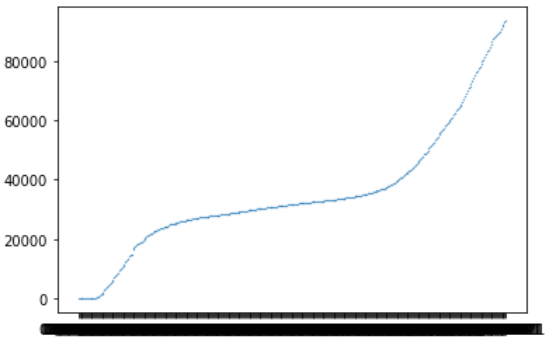

```
# positive number overtime
positiven_list = covid1['New Positives'].tolist()
p2 = plt.scatter(date_list,positiven_list, s=0.1);
```



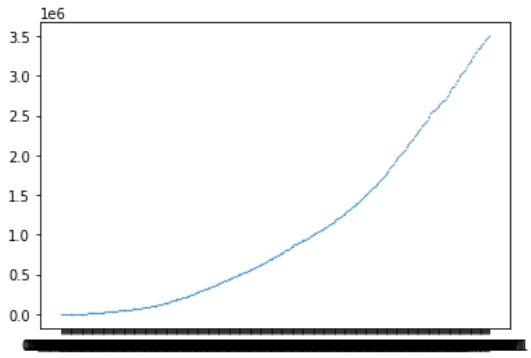

```
# total tests per day
totaltest_list = covid1['Total Number of Tests Performed'].tolist()
p3 = plt.scatter(date_list,totaltest_list,s=0.1);
```




```
# cumulative total positive test over time
cumtotalp_list = covid1['Cumulative Number of Positives'].tolist()
p4 = plt.scatter(date_list,cumtotalp_list, s=0.1);
```



```
# cumulative total test over time
cumtest_list = covid1['Cumulative Number of Tests Performed'].tolist()
p5 = plt.scatter(date_list,cumtest_list, s=0.1)
```



Correlations

```
covid1.corr()
```

	New Positives	Cumulative Number of Positives	Total Number of Tests Performed	Cumulative Number of Tests Performed	Positive Rate
New Positives	1.000000	0.560988	0.552787	0.617369	0.366361
Cumulative Number of Positives	0.560988	1.000000	0.830142	0.942585	-0.455666
Total Number of Tests Performed	0.552787	0.830142	1.000000	0.889318	-0.427358