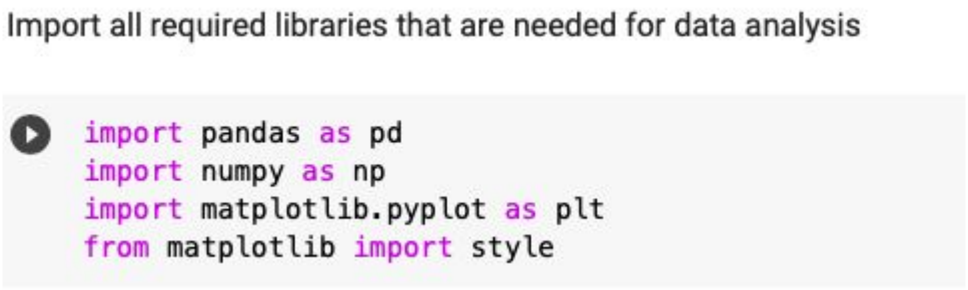
COVID 19 Analysis

Group member: **Kenneth Bentley, Xiaona Zhou, Tamzid Chowdhury, Vitoria Tai**

Task 1

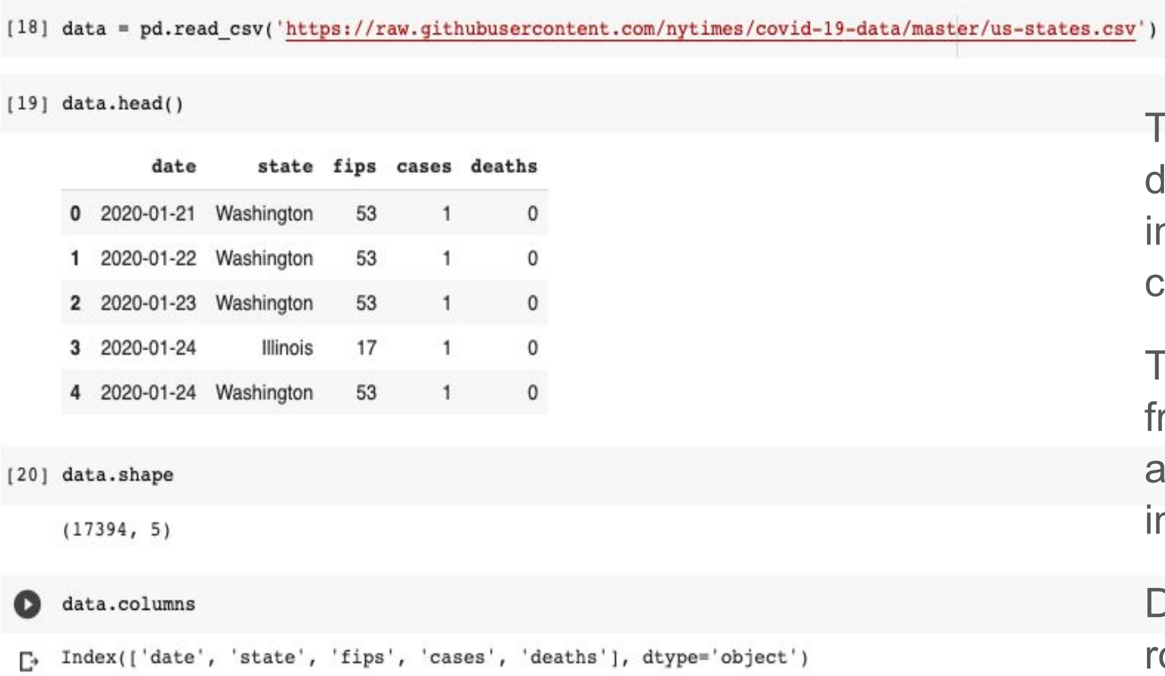
Required libraries:



Import all required libraries that are needed for data analysis

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from matplotlib import style
```

Task 2: Data Collection



```
[18] data = pd.read_csv('https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-states.csv')
```

```
[19] data.head()
```

	date	state	fips	cases	deaths
0	2020-01-21	Washington	53	1	0
1	2020-01-22	Washington	53	1	0
2	2020-01-23	Washington	53	1	0
3	2020-01-24	Illinois	17	1	0
4	2020-01-24	Washington	53	1	0

```
[20] data.shape

(17394, 5)
```

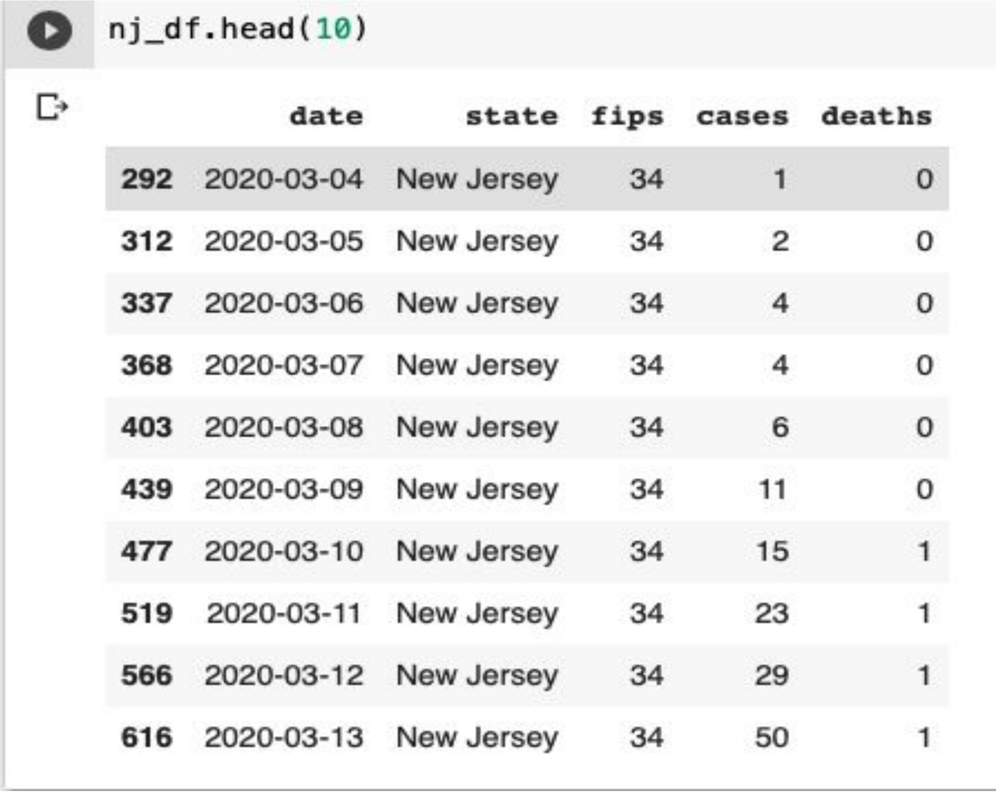
```
data.columns

Index(['date', 'state', 'fips', 'cases', 'deaths'], dtype='object')
```

The above link in red is the raw data from NYTimes on COVID-19 information on all the states in the country.

To help visualize the data in a data frame format, using data.head() allows us to see the first 5 cases in the list.

Data.shape shows the number of rows and columns being analized and data.columns() gives a clear idea of the topics being dealt with.

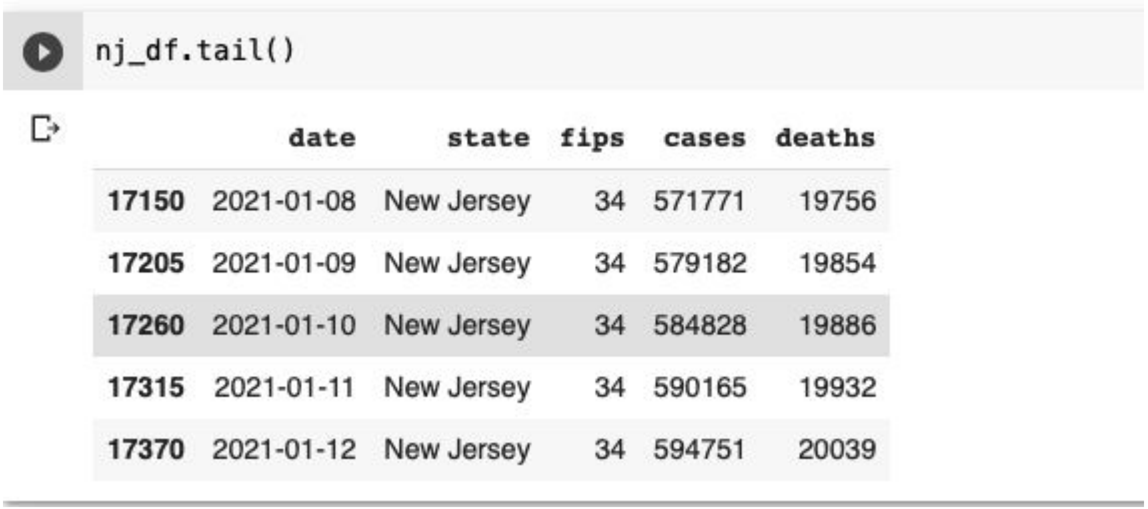
Task 3: Data Wrangling and EDA (Exploratory Data Analysis)



```
nj_df.head(10)
```

```
           date       state  fips  cases  deaths
292  2020-03-04  New Jersey    34      1       0
312  2020-03-05  New Jersey    34      2       0
337  2020-03-06  New Jersey    34      4       0
368  2020-03-07  New Jersey    34      4       0
403  2020-03-08  New Jersey    34      6       0
439  2020-03-09  New Jersey    34     11       0
477  2020-03-10  New Jersey    34     15       1
519  2020-03-11  New Jersey    34     23       1
566  2020-03-12  New Jersey    34     29       1
616  2020-03-13  New Jersey    34     50       1
```

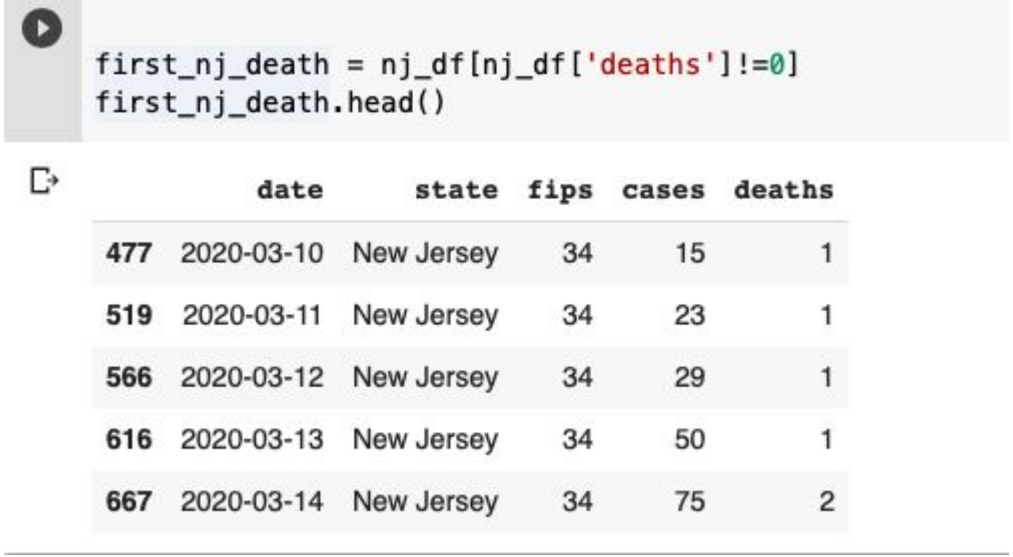
In the chart we can see the first ten days of Covid cases recorded in the state of New Jersey. This helps us understand how fast the virus started to spread and the cases started to rise as more and more people started to get infected.



```
nj_df.tail()
```

	date	state	fips	cases	deaths
17150	2021-01-08	New Jersey	34	571771	19756
17205	2021-01-09	New Jersey	34	579182	19854
17260	2021-01-10	New Jersey	34	584828	19886
17315	2021-01-11	New Jersey	34	590165	19932
17370	2021-01-12	New Jersey	34	594751	20039

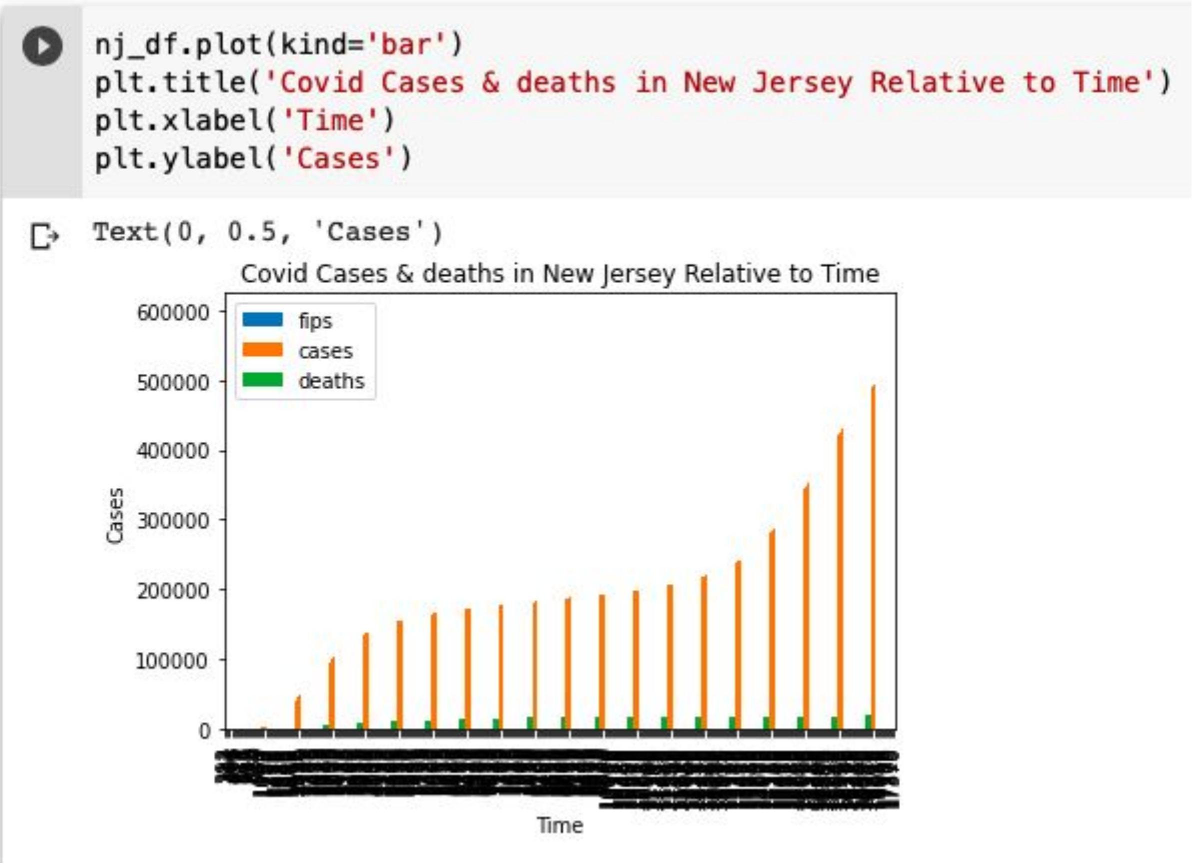
The chart above shows Covid cases recorded in the most recent days. As we can see from the previous slide, compared to the first week, the cases are increasing at an unbelievable and horrifying rate now. Compared to April 2nd of last year, exactly 1 month after the first reported case, to January 12th of this year, cases has increased almost 2324%.



```
first_nj_death = nj_df[nj_df['deaths']!=0]
first_nj_death.head()
```

	date	state	fips	cases	deaths
477	2020-03-10	New Jersey	34	15	1
519	2020-03-11	New Jersey	34	23	1
566	2020-03-12	New Jersey	34	29	1
616	2020-03-13	New Jersey	34	50	1
667	2020-03-14	New Jersey	34	75	2

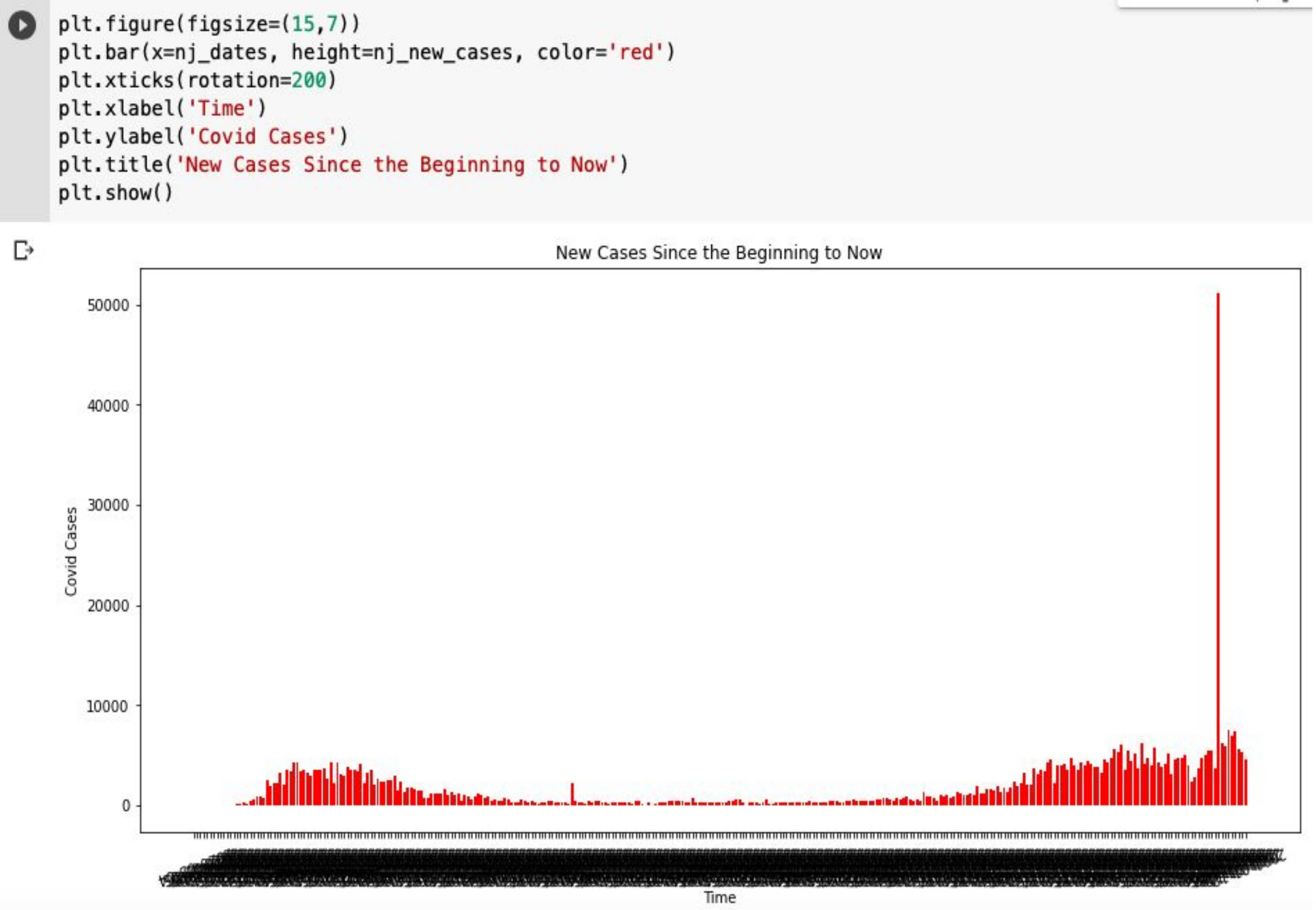
This chart shows the first death due to Covid in the state of New Jersey



```
nj_df.plot(kind='bar')
plt.title('Covid Cases & deaths in New Jersey Relative to Time')
plt.xlabel('Time')
plt.ylabel('Cases')
```

```
Text(0, 0.5, 'Cases')
```

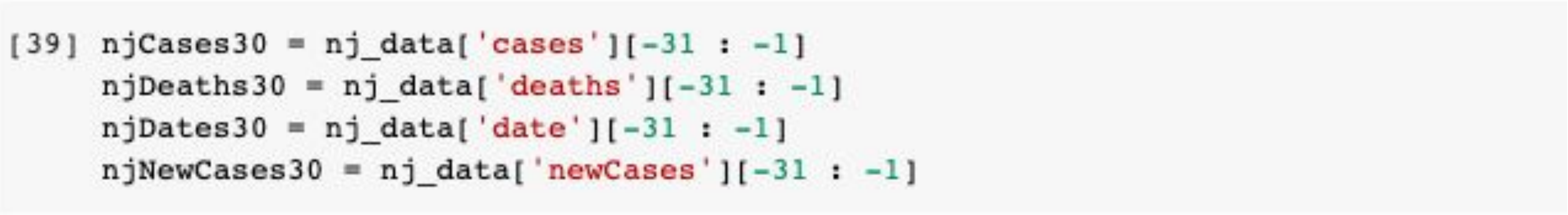
This bar graph shows the covid cases and deaths recorded relative to time in the state of New Jersey. Bar graph was a better choice to represent the data because we can see very easily how exponentially the cases grew as time passed since the first case.



```
plt.figure(figsize=(15,7))
plt.bar(x=nj_dates, height=nj_new_cases, color='red')
plt.xticks(rotation=200)
plt.xlabel('Time')
plt.ylabel('Covid Cases')
plt.title('New Cases Since the Beginning to Now')
plt.show()
```

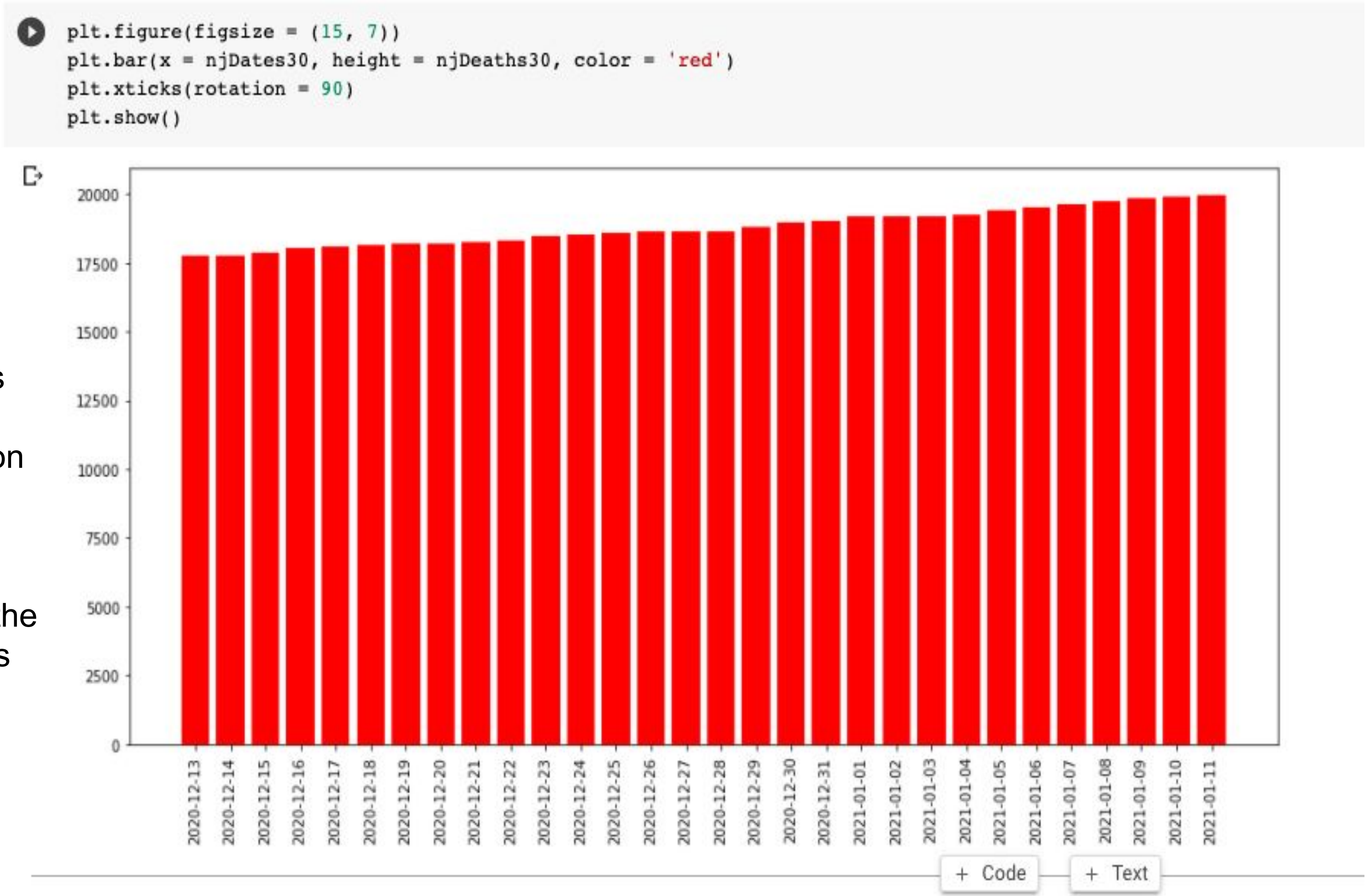
This chart shows the new cases recorded since the reporting of first case, to now. As we can see the cases started to slow down during the months of July - September, but it started to peak again since the last couple of months. And we've seen the highest case ever recorded in January of this year, when we see an abnormal spike in cases.

Task 4: Understand NJ COVID-19 data in the last 30 days



```
[39] njCases30 = nj_data['cases'][-31 : -1]
     njDeaths30 = nj_data['deaths'][-31 : -1]
     njDates30 = nj_data['date'][-31 : -1]
     njNewCases30 = nj_data['newCases'][-31 : -1]
```

Consolidating the last 30 days of cases, deaths, and new cases in New Jersey.

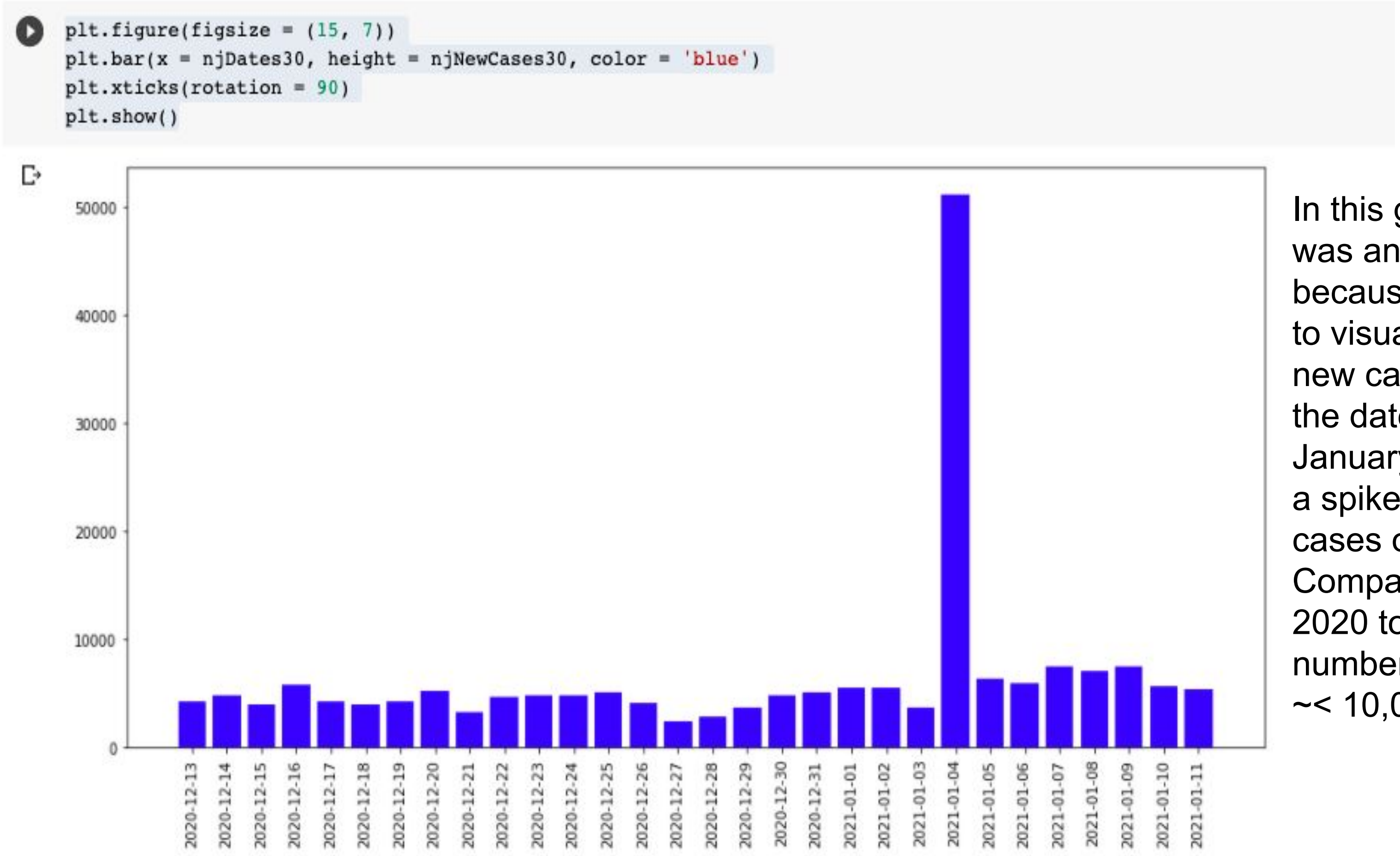


```
plt.figure(figsize = (15, 7))
plt.bar(x = njDates30, height = njDeaths30, color = 'red')
plt.xticks(rotation = 90)
plt.show()
```

In this graph, we can see the number of deaths from last 30 dates in New Jersey.

The histogram shows that the number of deaths was ~17500 on the first date and increasing steadily.

As of Jan. 11, 2021, the number of deaths has not passed 20,000



```
plt.figure(figsize = (15, 7))
plt.bar(x = njDates30, height = njNewCases30, color = 'blue')
plt.xticks(rotation = 90)
plt.show()
```

In this graph, a histogram was an appropriate choice because it makes it easier to visualize the number of new cases compared to the dates. As we can see, January 4, 2021, there was a spike of ~40,000 new cases on that day. Compared to Dec. 13, 2020 to jan 3, 2020, the numbers have been within ~< 10,000 new cases.

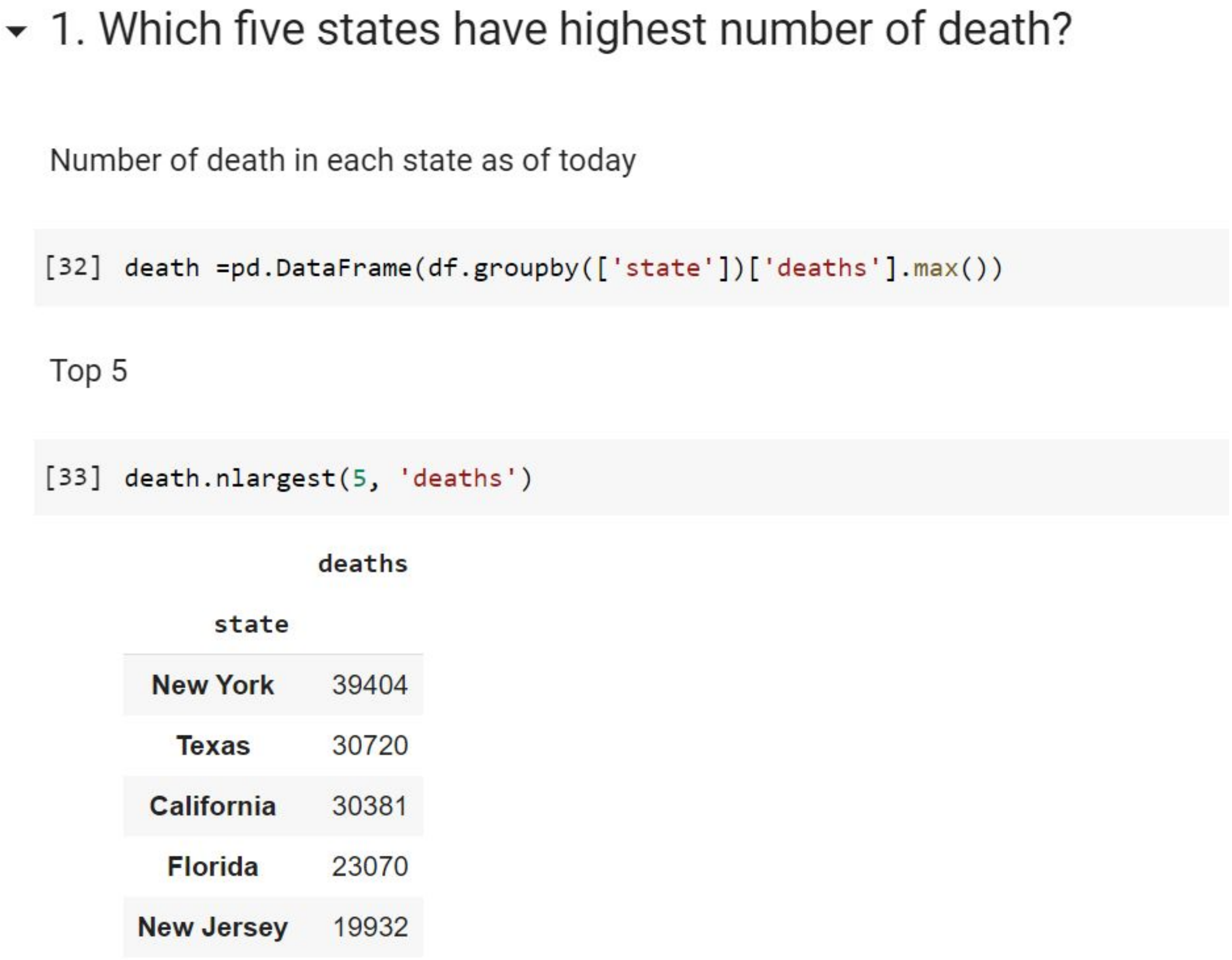
Conclusion

Based on our findings on the state of New Jersey:

1. The number of cases has been growing on a steady pace in New Jersey.
2. It has over 594,751 cases.
3. On January 4, 2021, the number of new cases spiked over 40,000.
4. Although New York and New Jersey are closed neighboring states, New York surpasses New Jersey on total cases by ~210,000.
5. Covid cases started to slow down during the months of July-September, but started to peak again during the month of November and has been increasing ever since.
6. Compared to April 2nd of last year, exactly 1 month after the first reported case, to January 12th of this year, the number of cases has increased by almost 2324%.

Extra work: Analysis on the states with highest number of death.

1. Which five states have the highest number of death?
2. Calculate new cases(different approach) and visualization
3. Calculate new deaths (different approach) and visualization



1. Which five states have highest number of death?

Number of death in each state as of today

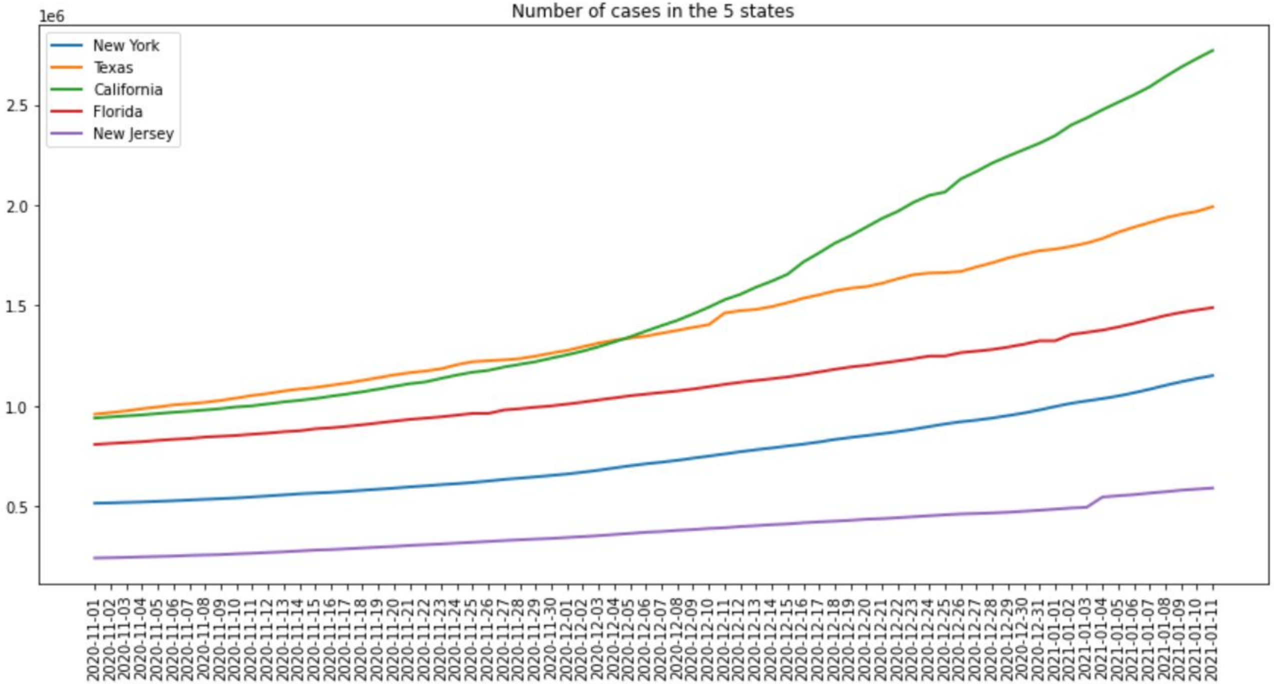
```
[32] death =pd.DataFrame(df.groupby(['state'])['deaths'].max())
```

Top 5

```
[33] death.nlargest(5, 'deaths')
```

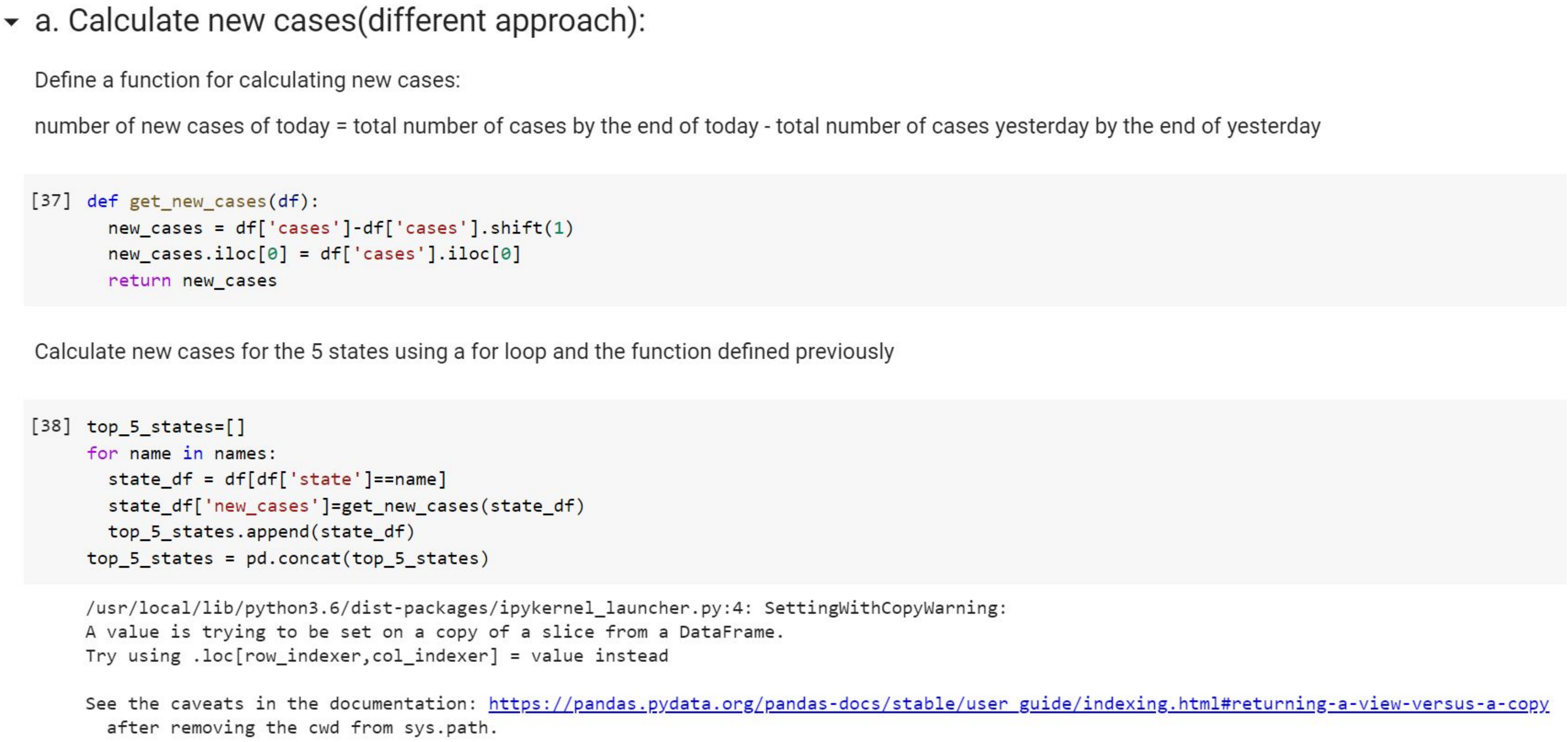
	deaths
state	
New York	39404
Texas	30720
California	30381
Florida	23070
New Jersey	19932

Visualize: number of cases in the five states



1. California surpass Texas on early December and become the worst in the nation.

2. The rate of increase was faster than the other states as well.



a. Calculate new cases(different approach):

Define a function for calculating new cases:

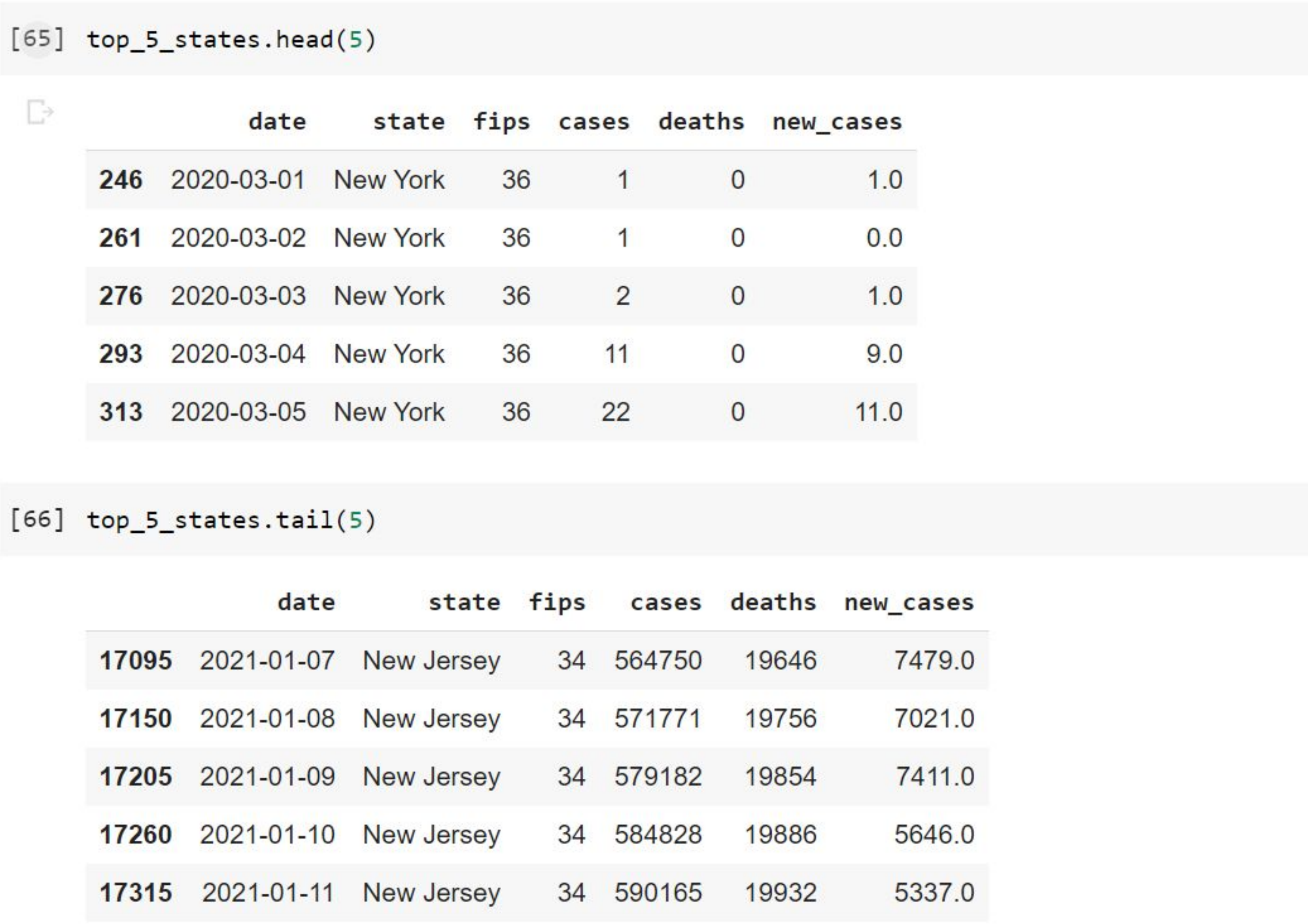
number of new cases of today = total number of cases by the end of today - total number of cases yesterday by the end of yesterday

```
[37] def get_new_cases(df):
       new_cases = df['cases']-df['cases'].shift(1)
       new_cases.iloc[0] = df['cases'].iloc[0]
       return new_cases
```

Calculate new cases for the 5 states using a for loop and the function defined previously

```
[38] top_5_states=[]
     for name in names:
       state_df = df[df['state']==name]
       state_df['new_cases']=get_new_cases(state_df)
       top_5_states.append(state_df)
     top_5_states = pd.concat(top_5_states)
```

```
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:4: 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
  after removing the cwd from sys.path.
```



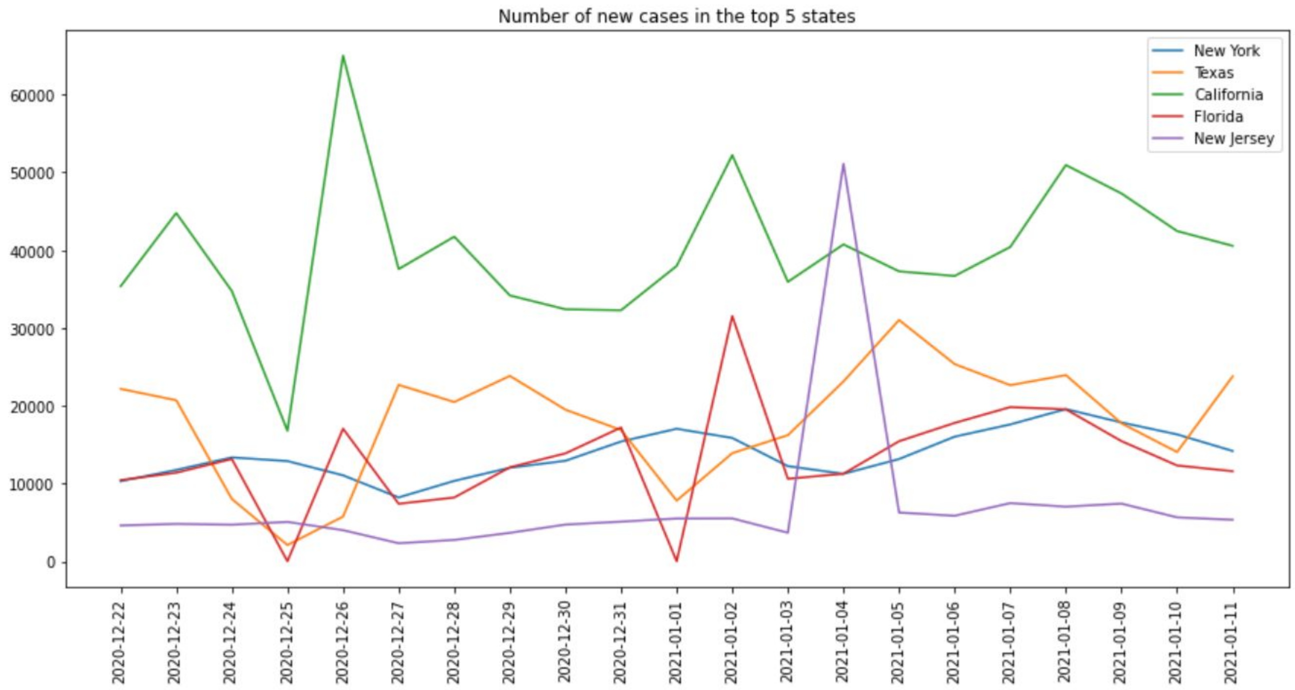
```
[65] top_5_states.head(5)
```

```
           date     state  fips  cases  deaths  new_cases
246  2020-03-01  New York    36      1       0        1.0
261  2020-03-02  New York    36      1       0        0.0
276  2020-03-03  New York    36      2       0        1.0
293  2020-03-04  New York    36     11       0        9.0
313  2020-03-05  New York    36     22       0       11.0
```

```
[66] top_5_states.tail(5)
```

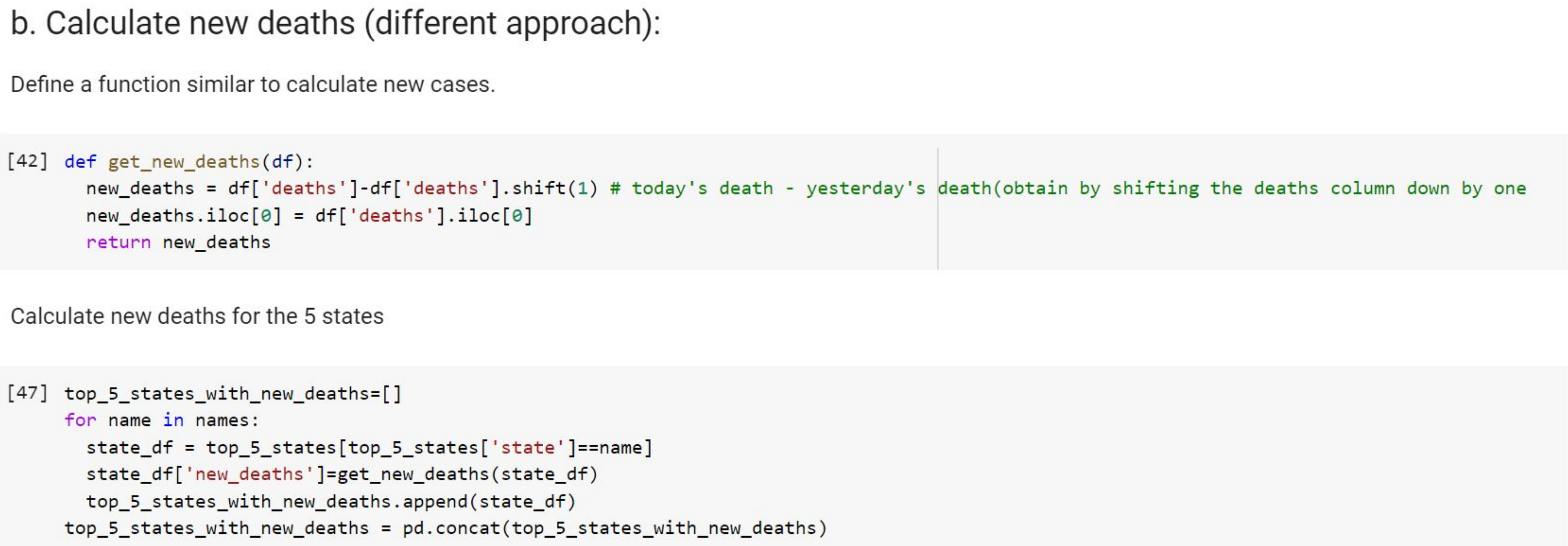
```
             date       state  fips   cases  deaths  new_cases
17095  2021-01-07  New Jersey    34  564750   19646     7479.0
17150  2021-01-08  New Jersey    34  571771   19756     7021.0
17205  2021-01-09  New Jersey    34  579182   19854     7411.0
17260  2021-01-10  New Jersey    34  584828   19886     5646.0
17315  2021-01-11  New Jersey    34  590165   19932     5337.0
```

Visualize: number of new cases in the five states



California had the highest number of new cases almost all the days we investigated (on January 4th, New Jersey had the highest number of new cases).

For New York, we see that the curve is quite flat but with noticeable increment, from about 10000 new cases a day to 20000 new cases a day.



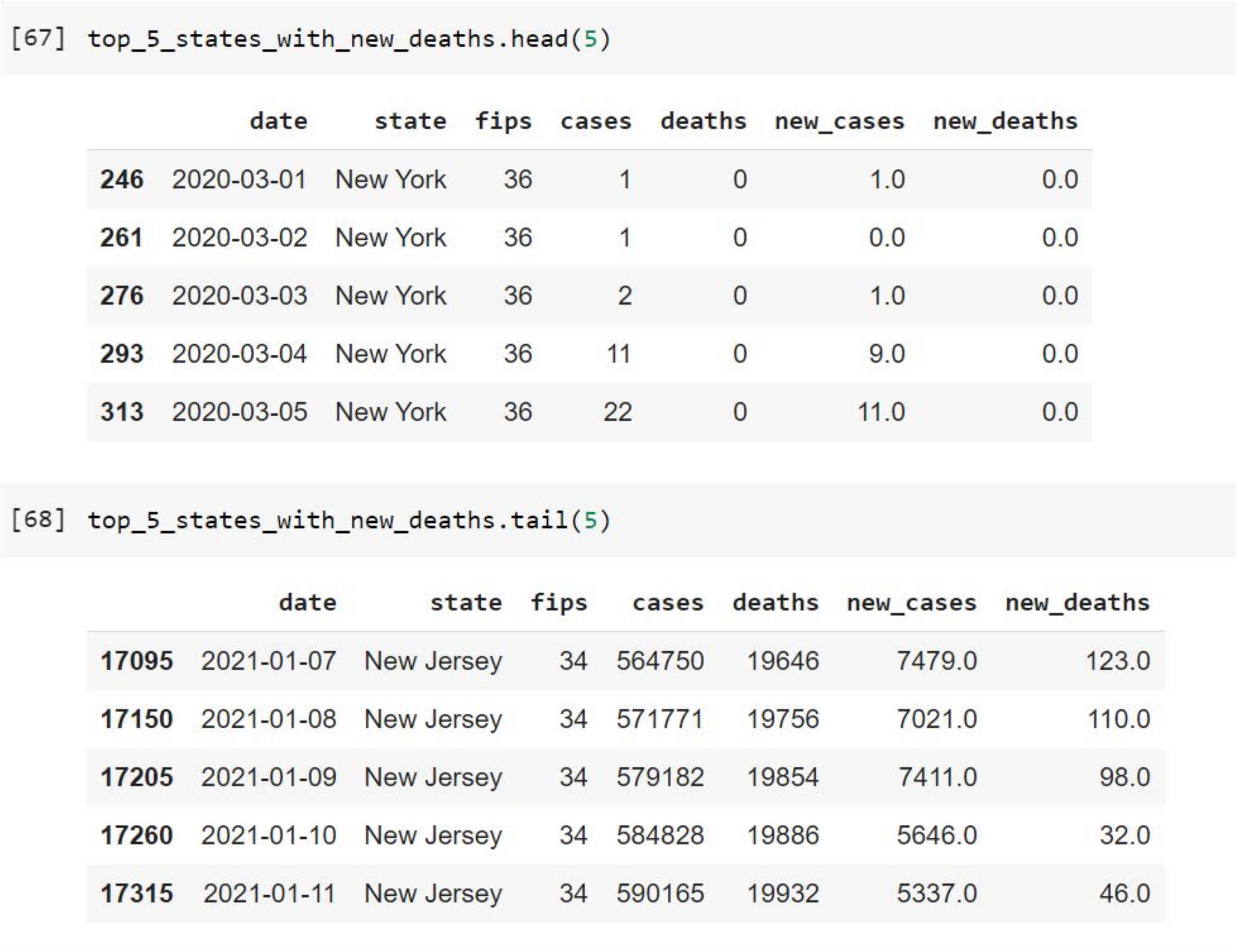
b. Calculate new deaths (different approach):

Define a function similar to calculate new cases.

```
[42] def get_new_deaths(df):
       new_deaths = df['deaths']-df['deaths'].shift(1) # today's death - yesterday's death(obtain by shifting the deaths column down by one
       new_deaths.iloc[0] = df['deaths'].iloc[0]
       return new_deaths
```

Calculate new deaths for the 5 states

```
[47] top_5_states_with_new_deaths=[]
     for name in names:
       state_df = top_5_states[top_5_states['state']==name]
       state_df['new_deaths']=get_new_deaths(state_df)
       top_5_states_with_new_deaths.append(state_df)
     top_5_states_with_new_deaths = pd.concat(top_5_states_with_new_deaths)
```



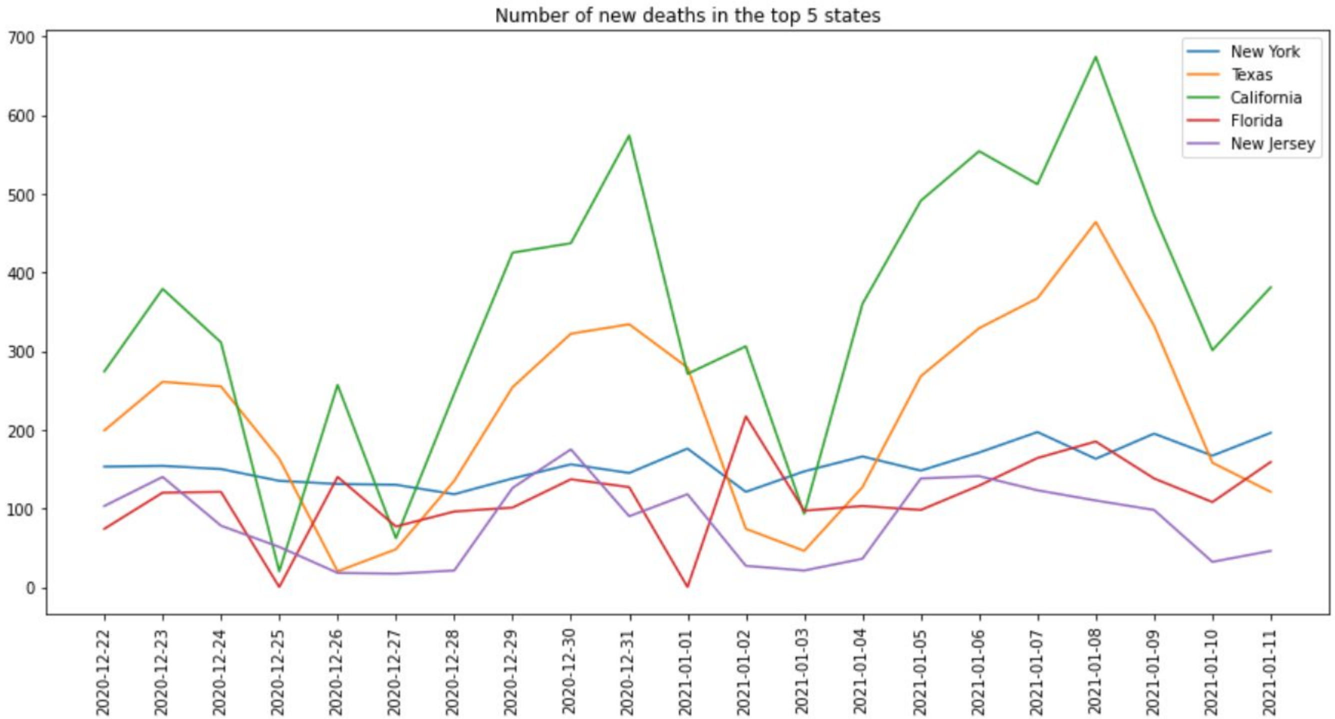
```
[67] top_5_states_with_new_deaths.head(5)
```

	date	state	fips	cases	deaths	new_cases	new_deaths
246	2020-03-01	New York	36	1	0	1.0	0.0
261	2020-03-02	New York	36	1	0	0.0	0.0
276	2020-03-03	New York	36	2	0	1.0	0.0
293	2020-03-04	New York	36	11	0	9.0	0.0
313	2020-03-05	New York	36	22	0	11.0	0.0

```
[68] top_5_states_with_new_deaths.tail(5)
```

	date	state	fips	cases	deaths	new_cases	new_deaths
17095	2021-01-07	New Jersey	34	564750	19646	7479.0	123.0
17150	2021-01-08	New Jersey	34	571771	19756	7021.0	110.0
17205	2021-01-09	New Jersey	34	579182	19854	7411.0	98.0
17260	2021-01-10	New Jersey	34	584828	19886	5646.0	32.0
17315	2021-01-11	New Jersey	34	590165	19932	5337.0	46.0

Visualize: number of new deaths in the five states



California had the highest number of deaths for most of the days we investigated. The maximum number of new deaths was 674 on 2021-01-08.

Every two minutes, one person dead in California on that day