```
import pandas as pd
df = pd.read_csv(r"D:\archive\covid_19_clean_complete.csv")
df
```

```
      Province/State         Country/Region        Lat       Long
Date  \
0                NaN            Afghanistan  33.939110  67.709953
2020-01-22
1                NaN                Albania  41.153300  20.168300
2020-01-22
2                NaN                Algeria  28.033900   1.659600
2020-01-22
3                NaN                Andorra  42.506300   1.521800
2020-01-22
4                NaN                 Angola -11.202700  17.873900
2020-01-22
...              ...                    ...        ...        ...
...
49063            NaN  Sao Tome and Principe   0.186400   6.613100
2020-07-27
49064            NaN                  Yemen  15.552727  48.516388
2020-07-27
49065            NaN                Comoros -11.645500  43.333300
2020-07-27
49066            NaN             Tajikistan  38.861000  71.276100
2020-07-27
49067            NaN                Lesotho -29.610000  28.233600
2020-07-27

       Confirmed  Deaths  Recovered  Active             WHO Region
0              0       0          0       0  Eastern Mediterranean
1              0       0          0       0                 Europe
2              0       0          0       0                 Africa
3              0       0          0       0                 Europe
4              0       0          0       0                 Africa
...          ...     ...        ...     ...                    ...
49063        865      14        734     117                 Africa
49064       1691     483        833     375  Eastern Mediterranean
49065        354       7        328      19                 Africa
49066       7235      60       6028    1147                 Europe
49067        505      12        128     365                 Africa

[49068 rows x 10 columns]
```

```
df =df.drop(columns=['Province/State'])

df.columns
```

```
Index(['Country/Region', 'Lat', 'Long', 'Date', 'Confirmed', 'Deaths',
       'Recovered', 'Active', 'WHO Region'],
      dtype='object')
```

```
df.dtypes
```

```
Country/Region     object
Lat               float64
Long              float64
Date               object
Confirmed           int64
Deaths              int64
Recovered           int64
Active              int64
WHO Region         object
dtype: object
```

```
df.duplicated()
```

```
0        False
1        False
2        False
3        False
4        False
         ...
49063    False
49064    False
49065    False
49066    False
49067    False
Length: 49068, dtype: bool
```

```
df.duplicated().sum()
```

```
np.int64(0)
```

```
df.duplicated(subset=['Country/Region', 'Date']).sum()
```

```
np.int64(13912)
```

```
df['Date'] = pd.to_datetime(df['Date'])
```

```
invalid_deaths = df[df['Deaths'] > df['Confirmed']]
```

```
invalid_recovered = df[df['Recovered'] > df['Confirmed']]
```

```
mismatch_active = df[df['Active'] != (df['Confirmed'] - (df['Deaths']
+ df['Recovered']))]
```

```
df = df.groupby(['Country/Region', 'Date'])[['Confirmed', 'Deaths',
'Recovered', 'Active']].sum().reset_index()
```

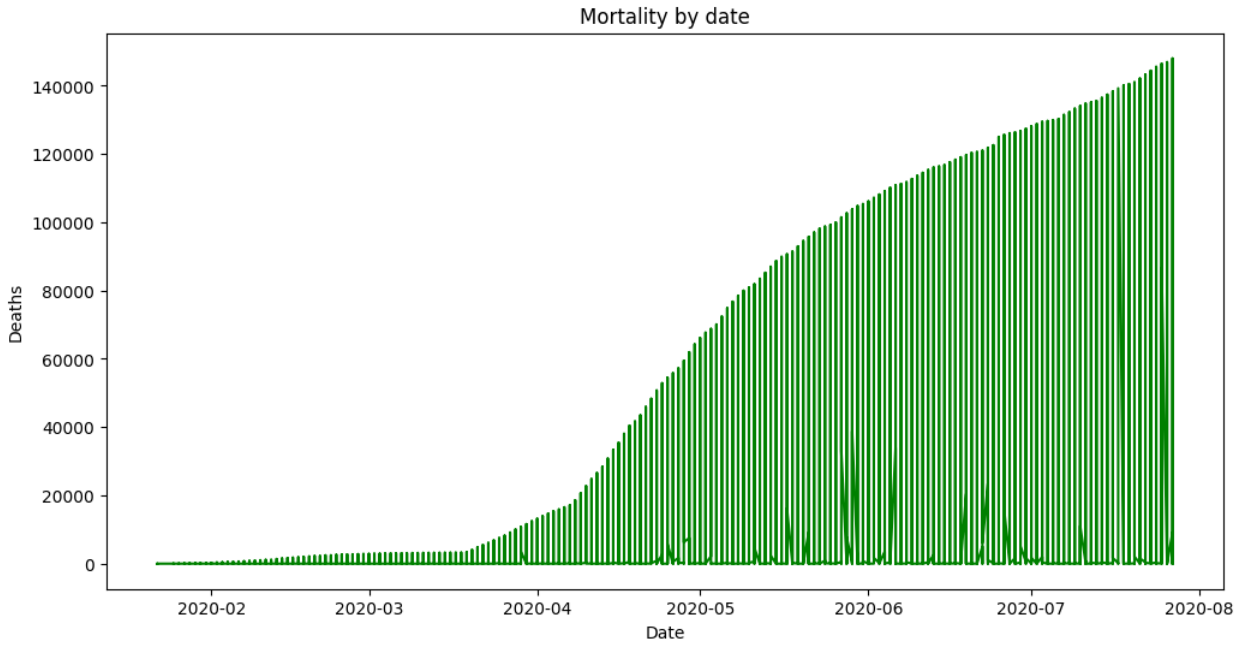
```
df['Date'] = pd.to_datetime(df['Date'],errors='coerce')

df.isnull().sum()
```

```
Country/Region    0
Date              0
Confirmed         0
Deaths            0
Recovered         0
Active            0
dtype: int64
```

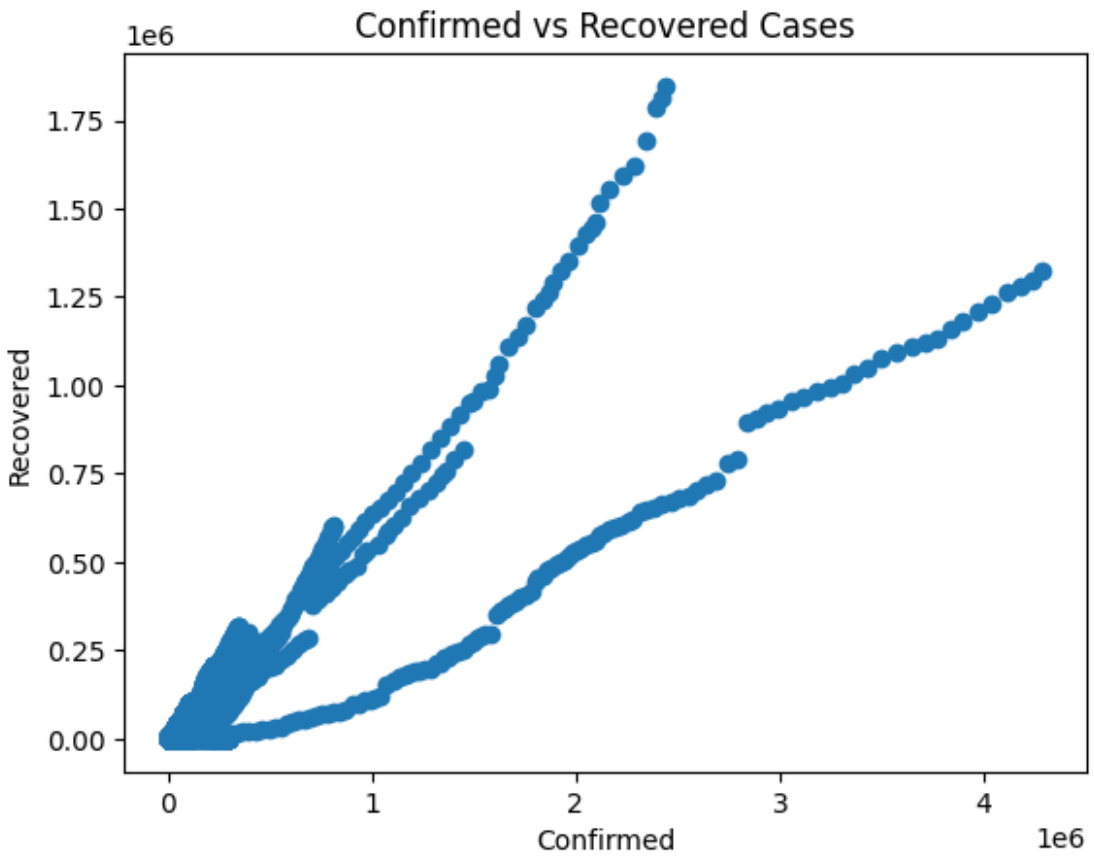
```
df.dtypes
```

```
Country/Region            object
Date              datetime64[ns]
Confirmed                  int64
Deaths                     int64
Recovered                  int64
Active                     int64
dtype: object
```

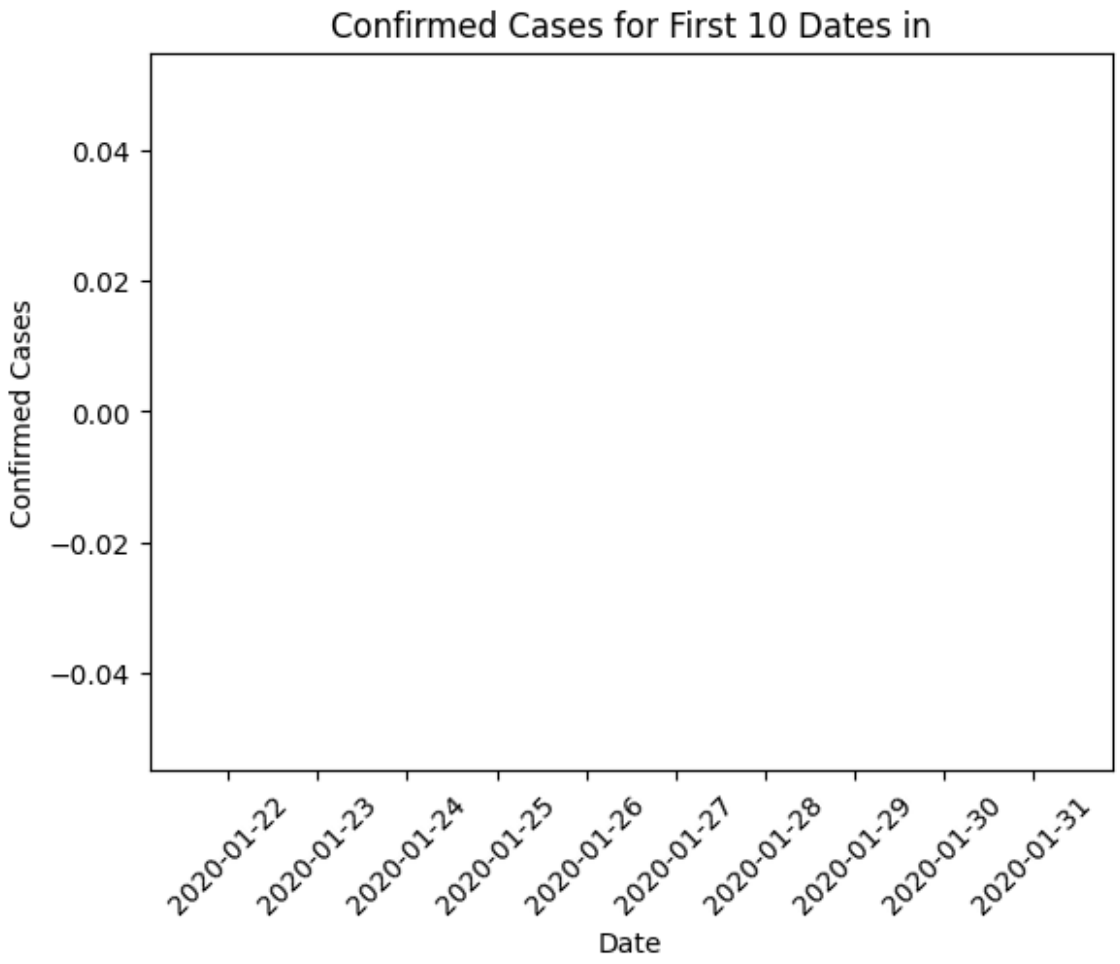
```
import matplotlib.pyplot as plt

df = df.sort_values('Date')

plt.figure(figsize=(12,6))
plt.plot(df['Date'],df['Deaths'],color='green')
plt.xlabel('Date')
plt.ylabel('Deaths')
plt.title('Mortality by date')
plt.show()
```



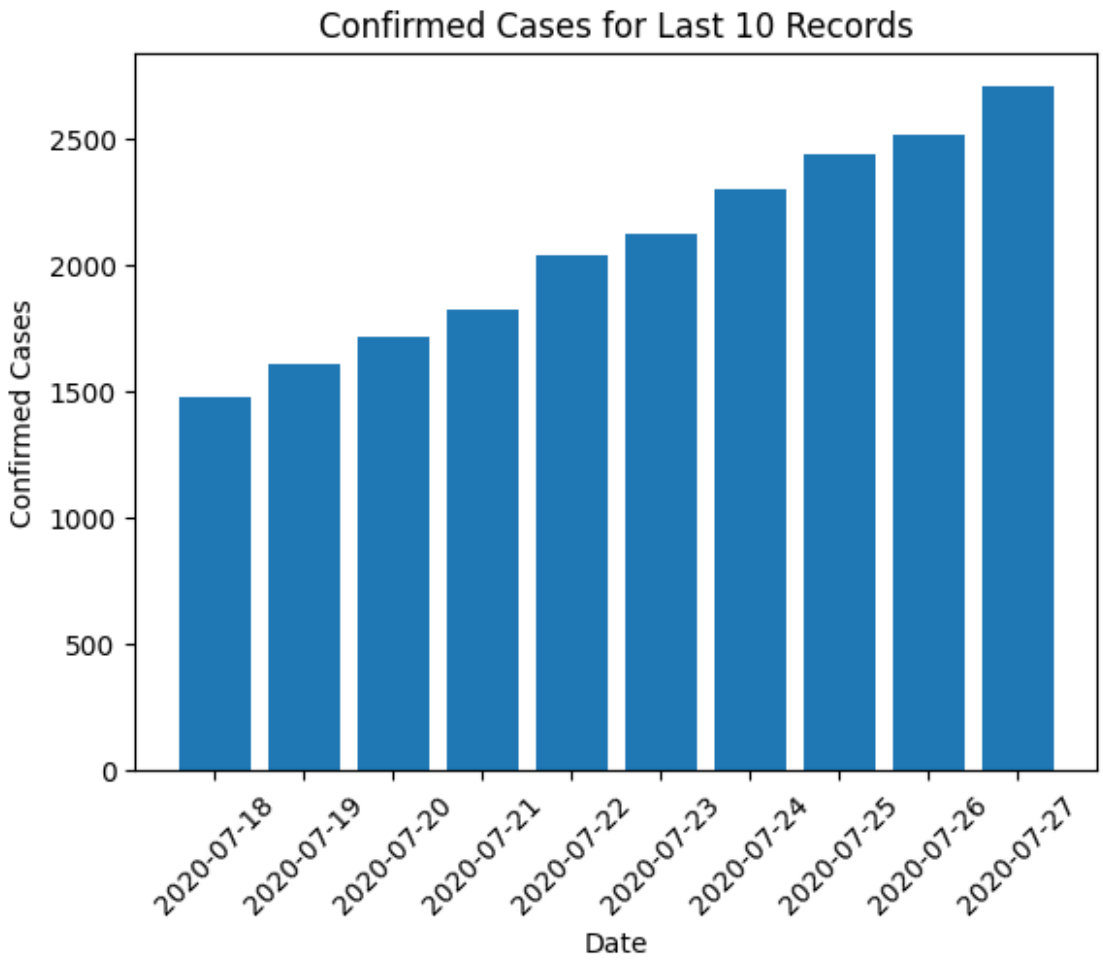

```
plt.scatter(df['Confirmed'],df['Recovered'])
plt.xlabel('Confirmed')
plt.ylabel('Recovered')
plt.title("Confirmed vs Recovered Cases")
plt.show()
```



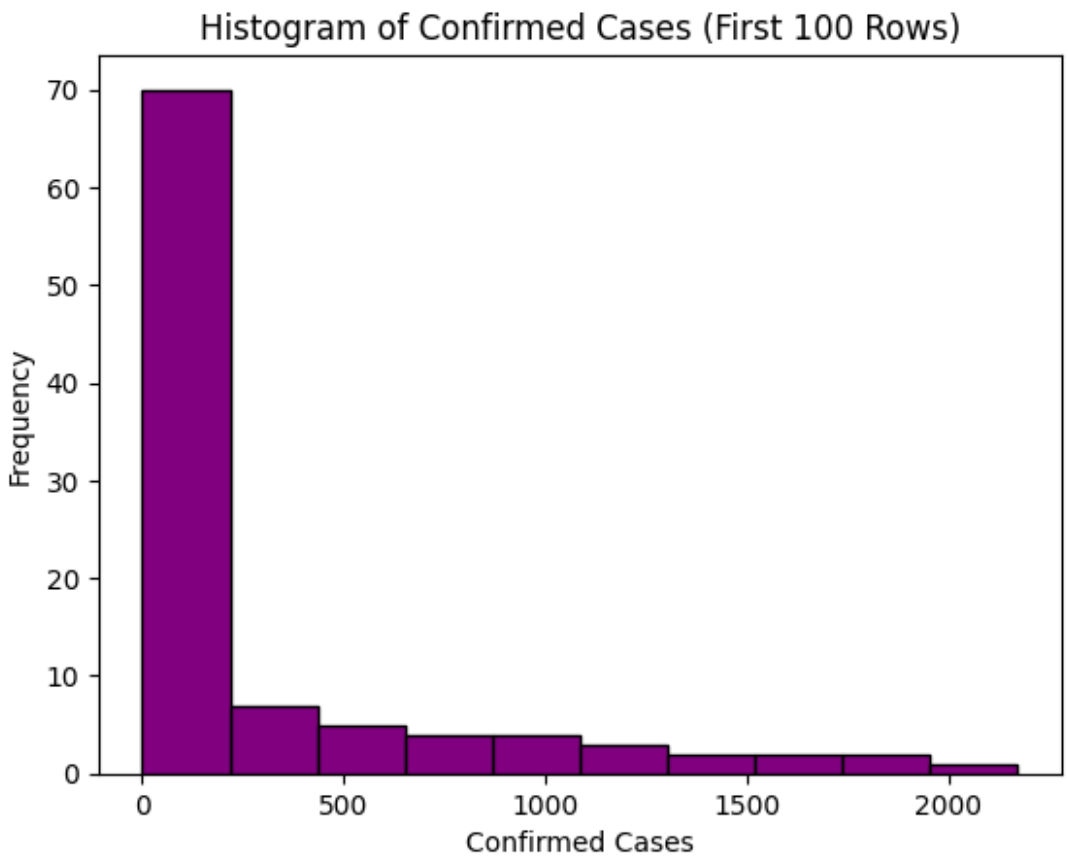

```
import matplotlib.pyplot as plt

plt.bar(df['Date'][:10], df['Confirmed'][:10])
plt.xlabel('Date')
plt.ylabel('Confirmed Cases')
plt.title('Confirmed Cases for First 10 Dates in')
plt.xticks(rotation=45)
plt.show()
```




```
import matplotlib.pyplot as plt

plt.bar(df['Date'].tail(10), df['Confirmed'].tail(10))
plt.xlabel('Date')
plt.ylabel('Confirmed Cases')
plt.title('Confirmed Cases for Last 10 Records')
plt.xticks(rotation=45)
plt.show()
```




```
import matplotlib.pyplot as plt


plt.hist(df['Confirmed'][:100], bins=10, color='purple',
edgecolor='black')

plt.xlabel('Confirmed Cases')
plt.ylabel('Frequency')
plt.title('Histogram of Confirmed Cases (First 100 Rows)')
plt.show()
```


Histogram of Confirmed Cases (First 100 Rows)
Frequency
Confirmed Cases
0
10
20
30
40
50
60
70
0
500
1000
1500
2000