

Unemployment Claims in the United States

Review the unemployment claims in the United State in 2019 and 2020.

Original data from <https://oui.doleta.gov/unemploy/claims.asp>

Import Dependencies

```
In [1]: import pandas as pd
import seaborn as sns
from matplotlib import pyplot as plt
import math
import numpy as np
import warnings

print('pandas version: ',pd.__version__)
print('seaborn version: ',sns.__version__)
print('numpy version: ',np.__version__)
```

```
pandas version: 1.2.4
seaborn version: 0.11.1
numpy version: 1.20.1
```

Grab Data

Read the csv file and store it. Display the first 5 rows as a preview.

```
In [2]: unemployment_data = pd.read_csv(r'./data/State_by_Year_Sorted.csv')

# Check the first few rows
unemployment_data.head()
```

```
Out[2]:
```

	State	Year	Claims	Claims_in_millions
0	California	2019	2099313	2.1
1	New York	2019	823926	0.8
2	Pennsylvania	2019	739399	0.7
3	Texas	2019	701860	0.7
4	New Jersey	2019	492540	0.5

Summarize the Numeric Columns

We can use `describe()` to summarize the numeric columns.

This will include the **Year** column, however, only the **Claims** column makes sense here.

```
In [3]: # Summarize the numeric columns
```

```
print('Total Row Count =', len(unemployment_data))
unemployment_data.describe()
```

Total Row Count = 106

Out[3]:

	Year	Claims	Claims_in_millions
count	106.000000	1.060000e+02	106.000000
mean	2019.500000	7.702800e+05	0.771698
std	0.502375	1.432701e+06	1.431134
min	2019.000000	1.532000e+03	0.000000
25%	2019.000000	1.010010e+05	0.100000
50%	2019.500000	2.830210e+05	0.300000
75%	2020.000000	8.377425e+05	0.800000
max	2020.000000	1.153777e+07	11.500000

Confirm Years in Data

Confirm that our data set only contains data for **2019** and **2020**.

In [4]:

```
# Confirm the number of years in the data.

data_years = unemployment_data['Year'].unique()
print('Years: ', data_years)

expected_number_of_years = 2
print('Number of years is 2? ', len(data_years) == 2)
```

Years: [2019 2020]
Number of years is 2? True

Summarize Claims Column

In [5]:

```
# Define the column that we'll be plotting
desired_column = 'Claims_in_millions'

# Summarize the Claims column
unemployment_data[desired_column].describe()
```

Out[5]:

count	106.000000
mean	0.771698
std	1.431134
min	0.000000
25%	0.100000
50%	0.300000
75%	0.800000
max	11.500000

Name: Claims_in_millions, dtype: float64

Exclude Outliers for Box Plot

Looking at the Claims data, we can assume some outliers beyond the lower 25% and upper 75%.

- 25% --> 0.1
- 75% --> 0.8

Let's round to get some cutoffs:

- Minimum --> 0
- Maximum --> 1

We'll also create two new data sets.

- 2019 --> unemployment_2019
- 2020 --> unemployment_2020

```
In [6]: # Exclude outliers
cutoff_min = 0
cutoff_max = 1
unemployment_data_limited = unemployment_data[unemployment_data[desired_column].
# Check the first few rows
unemployment_data_limited.head()
```

```
Out[6]:
```

	State	Year	Claims	Claims_in_millions
1	New York	2019	823926	0.8
2	Pennsylvania	2019	739399	0.7
3	Texas	2019	701860	0.7
4	New Jersey	2019	492540	0.5
5	Illinois	2019	489831	0.5

```
In [7]: # Create a collection of claims for each year (2019)
unemployment_2019 = unemployment_data_limited[(unemployment_data_limited.Year ==
# Check the first few rows of each (2019)
unemployment_2019.head()
```

```
Out[7]:
```

	State	Year	Claims	Claims_in_millions
1	New York	2019	823926	0.8
2	Pennsylvania	2019	739399	0.7
3	Texas	2019	701860	0.7
4	New Jersey	2019	492540	0.5
5	Illinois	2019	489831	0.5

```
In [8]: # Create a collection of claims for each year (2020)
unemployment_2020 = unemployment_data_limited[(unemployment_data_limited.Year ==
```

```
# Check the first few rows of each (2020)
unemployment_2020.head()
```

```
Out[8]:
```

	State	Year	Claims	Claims_in_millions
74	Missouri	2020	1014897	1.0
75	Tennessee	2020	981814	1.0
76	Oklahoma	2020	972453	1.0
77	Alabama	2020	917838	0.9
78	South Carolina	2020	842348	0.8

Describe 2019

```
In [9]: # Summarize the Claims column
unemployment_2019[desired_column].describe()
```

```
Out[9]: count    52.000000
mean      0.175000
std       0.188778
min       0.000000
25%      0.100000
50%      0.100000
75%      0.200000
max       0.800000
Name: Claims_in_millions, dtype: float64
```

Describe 2020

```
In [10]: # Summarize the Claims column
unemployment_2020[desired_column].describe()
```

```
Out[10]: count    32.000000
mean      0.440625
std       0.300386
min       0.000000
25%      0.200000
50%      0.300000
75%      0.700000
max       1.000000
Name: Claims_in_millions, dtype: float64
```

Monthly Line Chart

```
In [11]: # import matplotlib.pyplot as plt
# import pandas as pd

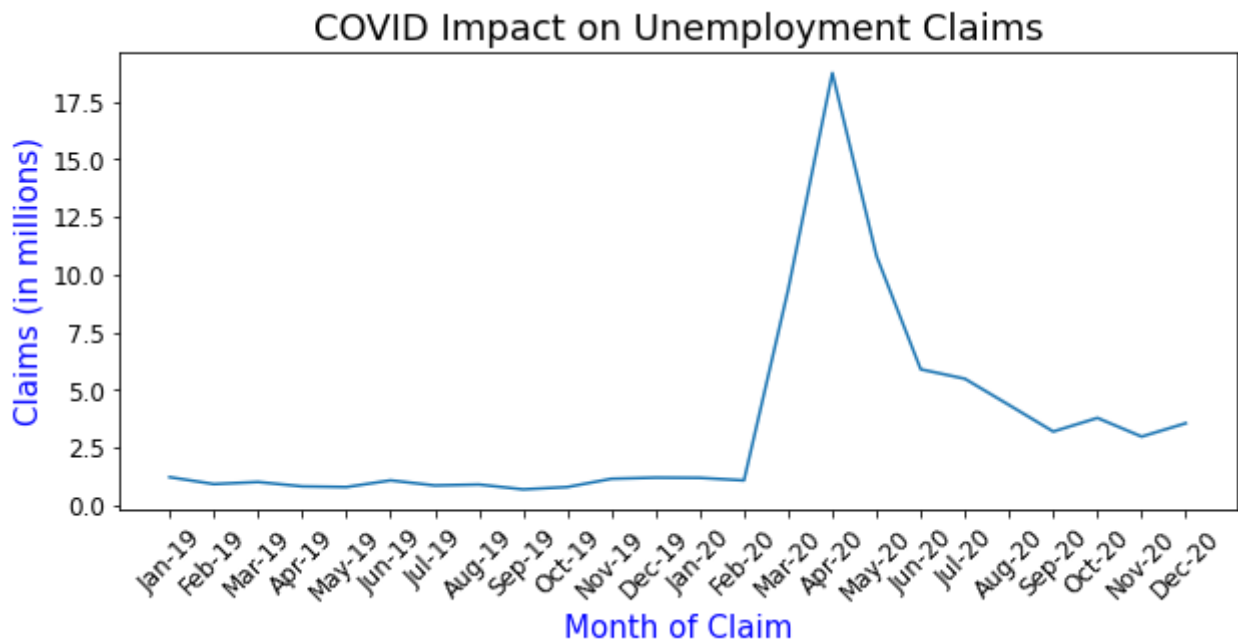
df = pd.read_csv(r'./data/Claims_by_Month.csv')

plt.figure(figsize=(10, 4.2))
x = range(len(df['Claims']))
plt.plot(x, df['Claims'])
```

```
plt.xticks(x, df['Month'])
plt.xticks(rotation = 45) # Rotates X-Axis Ticks by 45-degrees

plt.title('COVID Impact on Unemployment Claims', fontsize = 18)
plt.ylabel('Claims (in millions)', fontsize = 15, color = 'b')
plt.xlabel('Month of Claim', fontsize = 15, color = 'b')

plt.xticks(fontsize = 12)
plt.yticks(fontsize = 12)
plt.show()
```



Create Boxplots

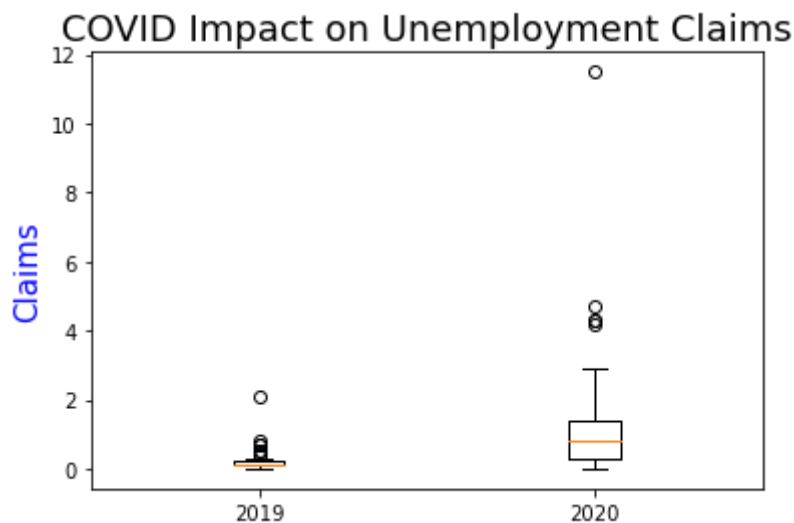
Create a boxplot with our new datasets, `unemployment_2019` and `unemployment_2020`.

```
In [12]: # Create a collection of claims for each year (2020)
unemployment_2020_full = unemployment_data[(unemployment_data.Year == 2020)]

# Create a collection of claims for each year (2020)
unemployment_2019_full = unemployment_data[(unemployment_data.Year == 2019)]

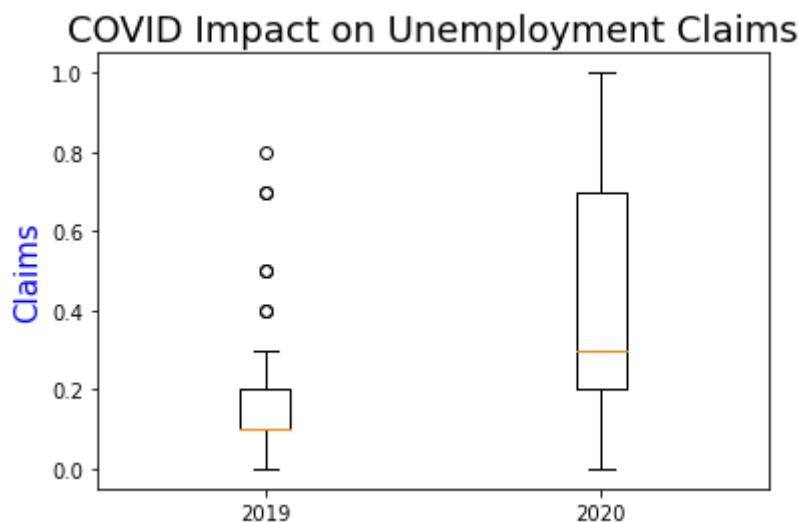
# Create a box plot for each year.
plt.boxplot([unemployment_2019_full[desired_column], unemployment_2020_full[desired_column]])
plt.ylabel('Claims', fontsize = 15, color = 'b')
plt.title('COVID Impact on Unemployment Claims', fontsize = 18)
```

```
Out[12]: Text(0.5, 1.0, 'COVID Impact on Unemployment Claims')
```



```
In [13]: # Create a box plot for each year.
plt.boxplot([unemployment_2019[desired_column], unemployment_2020[desired_column]]
plt.ylabel('Claims', fontsize = 15, color = 'b')
plt.title('COVID Impact on Unemployment Claims', fontsize = 18)
```

Out[13]: Text(0.5, 1.0, 'COVID Impact on Unemployment Claims')



Side-by-Side Histograms

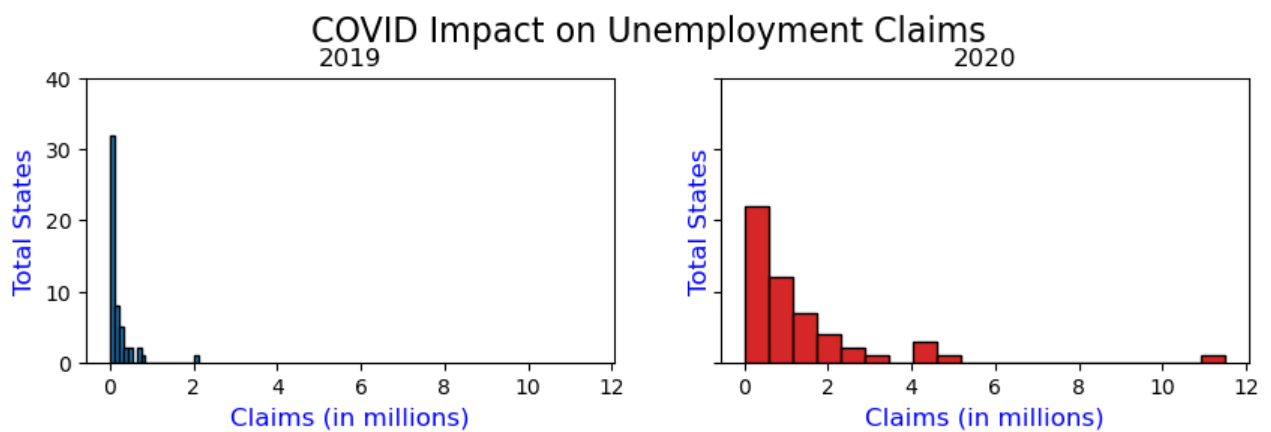
```
In [14]: fig, axes = plt.subplots(1, 2, figsize=(10,2.5), dpi=100, sharex=True, sharey=True)
colors = ['tab:blue', 'tab:red']

for i, (ax, Year) in enumerate(zip(axes.flatten(), unemployment_data.Year.unique)):
    x = unemployment_data.loc[unemployment_data.Year==Year, 'Claims_in_millions']
    ax.hist(x, bins=20, label=str(Year), color=colors[i], edgecolor = 'black')
    ax.set_title(Year)
    ax.set_xlabel('Claims (in millions)', fontsize = 12, color = 'b')
    ax.set_ylabel('Total States', fontsize = 12, color = 'b')

plt.ylim(0, 40)

plt.suptitle('COVID Impact on Unemployment Claims', y=1.05, size=16)
```

Out[14]: Text(0.5, 1.05, 'COVID Impact on Unemployment Claims')



Re-Employment Data

```
In [15]: df = pd.read_csv(r'./data/reemployment data/Combined Short_Sorted.csv')

fig, axes = plt.subplots(1, 2, figsize=(10,2.5), dpi=100, sharex=True, sharey=True)
colors = ['tab:blue', 'tab:red']

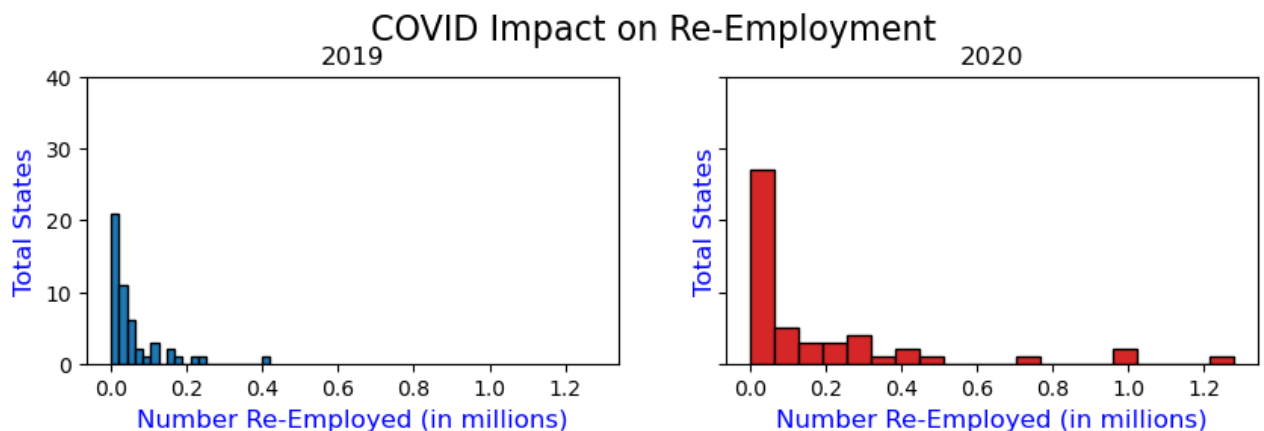
desired_column='# Remployed by next Quarter (millions)'
# desired_column='# Received UI Payment'

for i, (ax, Year) in enumerate(zip(axes.flatten(), df.Year.unique())):
    x = df.loc[df.Year==Year, desired_column]
    ax.hist(x, bins=20, label=str(Year), color=colors[i], edgecolor = 'black')
    ax.set_title(Year)
    ax.set_xlabel('Number Re-Employed (in millions)', fontsize = 12, color = 'b')
    ax.set_ylabel('Total States', fontsize = 12, color = 'b')

plt.ylim(0, 40)

plt.suptitle('COVID Impact on Re-Employment', y=1.05, size=16)
```

Out[15]: Text(0.5, 1.05, 'COVID Impact on Re-Employment')



In []: