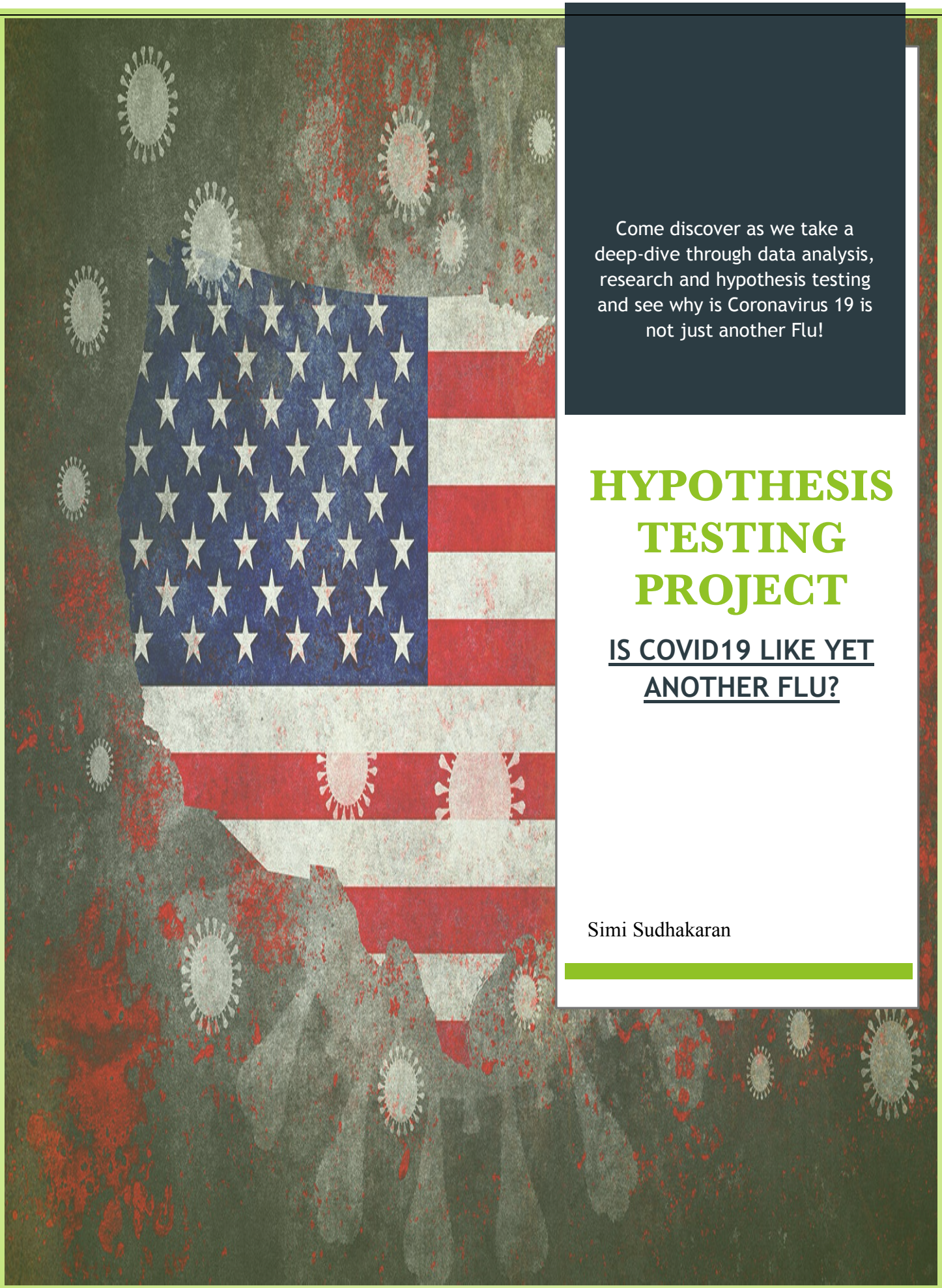


Come discover as we take a deep-dive through data analysis, research and hypothesis testing and see why is Coronavirus 19 is not just another Flu!

HYPOTHESIS TESTING PROJECT

IS COVID19 LIKE YET ANOTHER FLU?

Simi Sudhakaran

Table of Contents

INTRODUCTION

Since the new coronavirus was first discovered in January many people have compared it with a more well-known disease: The flu. Many of these comparisons pointed to the perhaps underappreciated toll of the flu. which causes millions of illnesses of tens of thousands of deaths every year in the U.S. alone (During the current flu season (Oct to Apr), the Centers for Disease Control and Prevention (CDC) estimates there were have been 39 million to 56 million flu illnesses and 24,000 to 62,000 flu death in the U.S., although that number is an estimate based on hospitalizations with flu symptoms, not based on actually counting every person who has died of flu)

The new coronavirus disease, COVID-19, has caused more than 1.2 million illnesses and 72,000 deaths in the U.S. as of May 6, 2020 according to data from John Hopkins University.

Both Covid-19 and the flu are respiratory illnesses. But Covid-19 is not the flu. Research so far indicates that Covid-19 spreads more easily and has a higher death rate than the flu,

Scientists are racing to find out more about COVID-19 and our understanding may change as new information becomes available. Based on what we know, we will compare Covid-19 and the Influenza(flu) cases and more importantly the death rate associated with the illnesses.

EXECUTIVE SUMMARY

Experts say there are a number of reasons why COVID-19 is a more serious illness than the seasonal flu. They point out there's no vaccine yet for COVID-19 and community-wide immunity hasn't built up. COVID-19 is also more infectious than the flu and has a higher death rate. COVID-19 also has a higher rate of hospitalizations. In this project herein, we will focus on comparing the death rate of Covid-19 and the Influenza on focus on rejecting the null hypothesis that the Covid-19 is just like any common flu. We will do this by the following:

Part 1: **Collection of the Dataset -** Data is the backbone of any sound decision making. In this part, we will focus on the data sources.

Part 2: **Exploratory Data Analysis (EDA)** - an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

Part 3: **Hypothesis Testing** - In theory, methods, and practice of testing a hypothesis by comparing it with the null hypothesis. The null hypothesis is only rejected if its probability falls below a predetermined significance level, in which case the hypothesis being tested is said to have that level of significance. In our case our Hypothesis Testing case would be:

Null Hypothesis (H_0): Covid19 is not any more dangerous than the normal flu (We will compare the death counts) Covid19 = Influenza

Alternate Hypothesis (H_a): Covid19 is more dangerous that the common influenza

Significance Level (α) = 0.05

We will calculate the test statistic and corresponding P-Value

Based on the results of our testing, we will form our conclusion and recommendations herein below.

TECHNOLOGY STACK

Python3 (Jupyter Notebook)

Python Codes

PART 1 (Collection of Datasets)

For this Project we will be using the below given datasets:

1. CDC Weekly Covid19 death counts from Feb'20 to May'20 in each state of US

   https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Week-Ending-D/r8kw-7aab/data

2. CDC Total Covid19 Death counts till May'20 in each state of US

   https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Sex-Age-and-S/9bhg-hcku/data

3. CDC Weekly Influenza Death counts in each state of US from 2019-2020

   https://gis.cdc.gov/grasp/fluview/mortality.html

4. NYTimes daily updated Covid19 death count in US

   https://raw.githubusercontent.com/nytimes/covid-19-data/master/us.csv

5. Abbreviation of each US state for geoMaps

   https://pe.usps.com/text/pub28/28apb.htm

PART 2 (Exploratory Data Analysis - EDA)

In any Data Science project, exploratory data analysis is the foundation in order to understand the project visually through graphs, plots, data tables, descriptive statistics etc.

For EDA we have used the below given libraries:

- Pandas
- Numpy
- Matplotlib
- Seaborn
- Plotly

The required datasets were also imported

Covid19 vs Influenza

In [65]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import plotly.express as px
import plotly.graph_objects as go

df_Covid_total = pd.read_excel("Provisional_COVID-19_Death_Counts_by_Sex__Age__and_State_IMP.xls")
df_flu_2019 = pd.read_csv("State_Custom_Data_IMP.csv")
df_Covid_weekly = pd.read_excel("weekly state-wise death by covid19 and flu 2020.xls")
```

In [66]:
```
df_Covid_total.head()
```

Out[66]:

	Data as of	Start week	End Week	State	Sex	Age group	COVID-19 Deaths	Total Deaths	Pneumonia Deaths	Pneumonia and COVID-19 Deaths	Influenza Deaths	Pneumonia, Influenza, or COVID-19 Deaths	Footnote
0	2020-05-06	2020-02-01	2020-05-02	United States	All Sexes	Under 1 year	4.0	3951.0	36.0	1.0	11.0	50.0	NaN
1	2020-05-06	2020-02-01	2020-05-02	United States	All Sexes	1-4 years	2.0	780.0	33.0	2.0	33.0	66.0	NaN
2	2020-05-06	2020-02-01	2020-05-02	United States	All Sexes	5-14 years	4.0	1146.0	38.0	0.0	41.0	83.0	NaN
3	2020-05-06	2020-02-01	2020-05-02	United States	All Sexes	15-24 years	48.0	6843.0	143.0	18.0	41.0	211.0	NaN
4	2020-05-06	2020-02-01	2020-05-02	United States	All Sexes	25-34 years	317.0	14629.0	496.0	134.0	133.0	800.0	NaN

In [67]:
```
df_flu_2019.head()
```

Out[67]:

	AREA	SUB AREA	AGE GROUP	SEASON	WEEK	PERCENT P&I	NUM INFLUENZA DEATHS	NUM PNEUMONIA DEATHS	TOTAL DEATHS	PERCENT COMPLETE
0	State	Alabama	All	2019-20	40	5.0	0	48	955	98.3%
1	State	Alabama	All	2019-20	41	3.8	0	36	940	96.8%
2	State	Alabama	All	2019-20	42	4.5	0	45	1,011	> 100%
3	State	Alabama	All	2019-20	43	4.4	2	42	1,000	> 100%

Preprocessing the Datasets

Here we are cleaning up the dataset for all the garbage/null values and dropping the irrelevant columns

```
In [69]: df_Covid_total.isna().sum()
```

```
Out[69]: Data as of                                  0
         Start week                                  0
         End Week                                    0
         State                                       0
         Sex                                         0
         Age group                                   0
         COVID-19 Deaths                           308
         Total Deaths                              193
         Pneumonia Deaths                          321
         Pneumonia and COVID-19 Deaths             334
         Influenza Deaths                          528
         Pneumonia, Influenza, or COVID-19 Deaths  348
         Footnote                                  489
         dtype: int64
```

Preprocessing the dataset

```
In [70]: df_Covid_total = df_Covid_total.drop(columns="Total Deaths")
         df_Covid_total = df_Covid_total.drop(columns="Pneumonia Deaths")
         df_Covid_total = df_Covid_total.drop(columns="Pneumonia and COVID-19 Deaths")
         df_Covid_total = df_Covid_total.drop(columns="Pneumonia, Influenza, or COVID-19 Deaths")
         df_Covid_total = df_Covid_total.drop(columns="Footnote")
```

```
In [71]: df_Covid_total = df_Covid_total.rename(columns=lambda x: x.lower().replace(' ', '_'))
         df_Covid_total.head()
```

Out[71]:

	data_as_of	start_week	end_week	state	sex	age_group	covid-19_deaths	influenza_deaths
0	2020-05-06	2020-02-01	2020-05-02	United States	All Sexes	Under 1 year	4.0	11.0

After that we have filtered the dataset row-wise. We took all the age group total and all the sex totals for the Covid19 and Influenza fatality in each state of US and stored it in a dataframe

```
In [72]: filteredSex_df_Covid_total = df_Covid_total.sex.isin(["All sexes"])
         filteredAge_df_Covid_total = df_Covid_total.age_group.isin(["All Ages"])
         df_Covid_total[filteredSex_df_Covid_total]
         df_Covid_total[filteredAge_df_Covid_total]
```

Out[72]:

	data_as_of	start_week	end_week	state	sex	age_group	covid-19_deaths	influenza_deaths
11	2020-05-06	2020-02-01	2020-05-02	United States	All Sexes Total	All Ages	44016.0	5971.0
37	2020-05-06	2020-02-01	2020-05-02	United States Total	All sexes	All Ages	44016.0	5971.0
63	2020-05-06	2020-02-01	2020-05-02	Alabama Total	All sexes	All Ages	208.0	85.0
89	2020-05-06	2020-02-01	2020-05-02	Alaska Total	All sexes	All Ages	NaN	NaN
115	2020-05-06	2020-02-01	2020-05-02	Arizona Total	All sexes	All Ages	294.0	106.0
141	2020-05-06	2020-02-01	2020-05-02	Arkansas Total	All sexes	All Ages	44.0	67.0
167	2020-05-06	2020-02-01	2020-05-02	California Total	All sexes	All Ages	1419.0	555.0
193	2020-05-06	2020-02-01	2020-05-02	Colorado Total	All sexes	All Ages	627.0	92.0

5-Point Summary

Thereafter, we have calculated the 5-point summary which includes:

- Mean
- Median
- Maximum value
- Minimum value
- Quartile 1
- Quartile 3

of the Covid19 and Influenza data.

5-point summary of just 3.2 months of Covid19 fatality, clearly proves that the rate of progression and mortality with Covid19 is far higher in than 1complete year of Influenza deaths in US.

5 Point Summary of the Covid19 death(Feb'20-May'20) vs Influenza deaths(2020-2019)

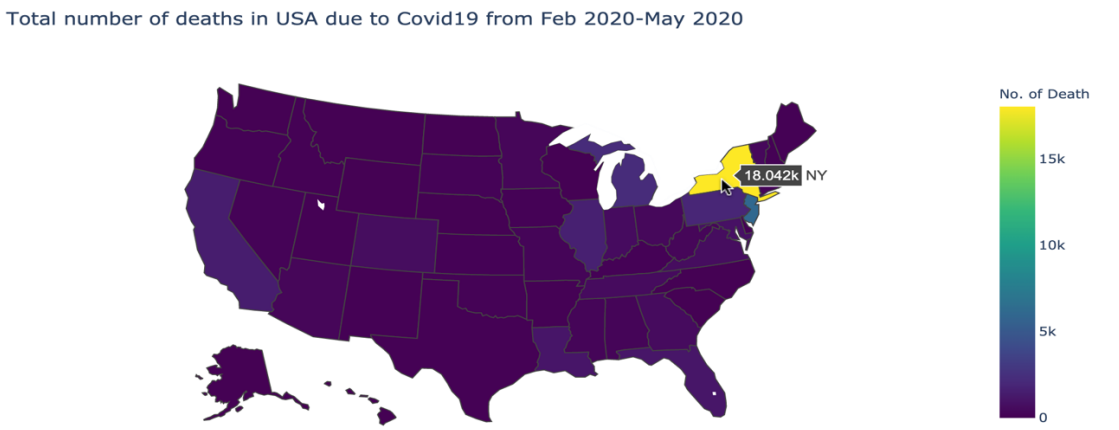
In [60]:
```
print("Mean of Covid19 Fatalities from Feb 2020 - May 2020: ", dfFinal_Covid_total.loc[:,"covid-19_deaths"].mean())
print("Mean of Influenza Fatalities in 2019-2020: ", dfFinal_Covid_total.loc[:,"influenza_deaths"].mean())
print("Median of Covid19 Fatalities Feb 2020 - May 2020: ", dfFinal_Covid_total.loc[:,"covid-19_deaths"].median())
print("Median of Influenza Fatalities in 2019-2020: ", dfFinal_Covid_total.loc[:,"influenza_deaths"].median())
print("Maximum value of Covid19 Fatalities Feb 2020 - May 2020: ", np.nanmax(dfFinal_Covid_total.loc[:,"covid-19_deaths
print("Maximum value of Influenza Fatalities in 2019-2020: ", np.nanmax(dfFinal_Covid_total.loc[:,"influenza_deaths"].v
print("Minimum value of Covid19 Fatalities Feb 2020 - May 2020: ", np.nanmin(dfFinal_Covid_total.loc[:,"covid-19_deaths
print("Minimum value of Influenza Fatalities in 2019-2020: ", np.nanmin(dfFinal_Covid_total.loc[:,"influenza_deaths"].v
print("Q1 of Covid19 Fatalities Feb 2020 - May 2020: ", dfFinal_Covid_total.loc[:,"covid-19_deaths"].quantile(0.25))
print("Q1 of Influenza Fatalities in 2019-2020: ", dfFinal_Covid_total.loc[:,"influenza_deaths"].quantile(0.25))
print("Q3 of Covid19 Fatalities Feb 2020 - May 2020: ", dfFinal_Covid_total.loc[:,"covid-19_deaths"].quantile(0.75))
print("Q3 of Influenza Fatalities in 2019-2020: ", dfFinal_Covid_total.loc[:,"influenza_deaths"].quantile(0.75))
```

```
Mean of Covid19 Fatalities from Feb 2020 - May 2020:  1067.9622641509434
Mean of Influenza Fatalities in 2019-2020:  129.0754716981132
Median of Covid19 Fatalities Feb 2020 - May 2020:  169.0
Median of Influenza Fatalities in 2019-2020:  85.0
Maximum value of Covid19 Fatalities Feb 2020 - May 2020:  18042.0
Maximum value of Influenza Fatalities in 2019-2020:  1081.0
Minimum value of Covid19 Fatalities Feb 2020 - May 2020:  0.0
Minimum value of Influenza Fatalities in 2019-2020:  0.0
Q1 of Covid19 Fatalities Feb 2020 - May 2020:  48.0
Q1 of Influenza Fatalities in 2019-2020:  27.0
Q3 of Covid19 Fatalities Feb 2020 - May 2020:  585.0
Q3 of Influenza Fatalities in 2019-2020:  122.0
```

Heatmap of Covid19 deaths in the United States (state-wise)

This heat map shows the total number of deaths due to Covid19 in all the states of US from Feb'20 to May'20. This is a hover on heat map which would display the total deaths in that particular state. The intensity of the fatalities is represented by the heat legend. Here we can clearly see that New York was the worst affected with close to 27000 deaths till May 12, 2020.

```
fig = go.Figure(data=go.Choropleth(
    locations=df_usa_locAbb['Abbreviation'], # Spatial coordinates
    z = df_2['covid-19_deaths'], # Data to be color-coded
    locationmode = 'USA-states', # set of locations match entries in `locations`
    colorscale = 'Viridis',
    colorbar_title = "No. of Death",
))

fig.update_layout(
    title_text = 'Total number of deaths in USA due to Covid19 from Feb 2020-May 2020',
    geo_scope='usa', # limit map scope to USA
)

fig.show()
```



This is a table displaying the 5 worst affected states in US due to Covid19

```
In [88]: fig = go.Figure(data=[go.Table(
             header=dict(values=[['<b>STATE</b>'],['<b>COVID19 DEATHS</b>']],
                         fill_color='paleturquoise',
                         align='center'),
             cells=dict(values=[df_3.state, df_3['covid-19_deaths']],
                        fill_color='lavender',
                        align='center'))
         ])

         fig.show()
```

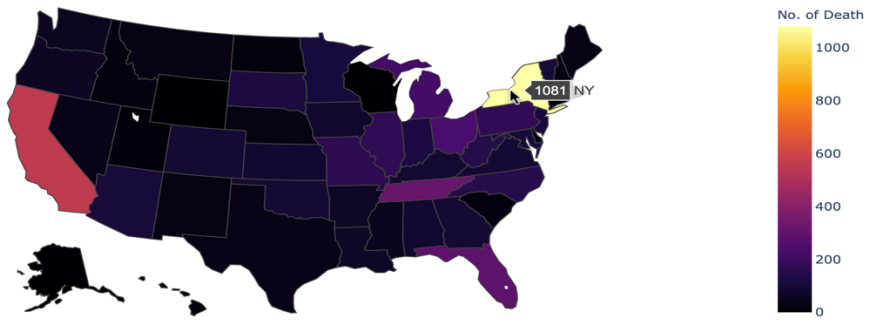
STATE	COVID19 DEATH
New York Total	18042
New Jersey Total	5991
Massachusetts Total	2752
Michigan Total	2238
Pennsylvania Total	1908

Heat Map of the Influenza deaths in the United States (state-wise)

This heatmap shows the total number of deaths due to Influenza in US state-wise from Feb 2020 to May 2020. This is a hover on heatmap which displays the total deaths due to Influenza in that particular state. Also, the color code intensity is in accordance with the number of fatalities. There are a few states with blanks because the data wasn't available for those states.

Below we can clearly see that there were 1081 fatalities in NY till May'20 for the year 2020.

```
In [122]:
fig = go.Figure(data=go.Choropleth(
    locations=df_usa_locAbb['Abbreviation'], # Spatial coordinates
    z = df_2['influenza_deaths'], # Data to be color-coded
    locationmode = 'USA-states', # set of locations match entries in `locations`
    colorscale = 'Inferno',
    colorbar_title = "No. of Death",
))

fig.update_layout(
    title_text = 'Total number of deaths in USA due to Influenza from Feb 2020-May2020',
    geo_scope='usa', # limit map scope to USA
)

fig.show()
```



The table below displays the 5 most affected states by Influenza

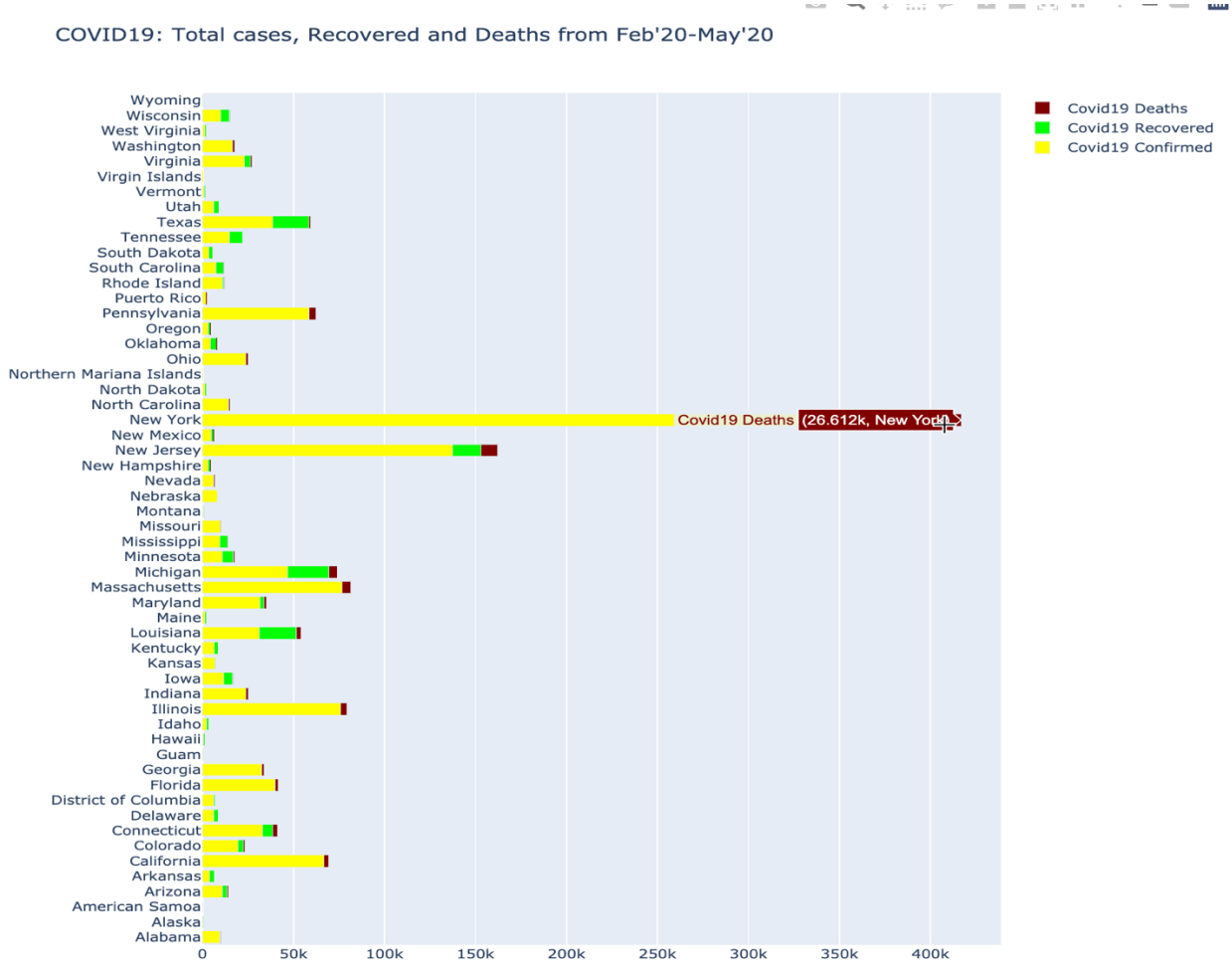
```
In [103]:
df_2 = df_1[["state","influenza_deaths"]]
df_3 = df_2.nlargest(5,"influenza_deaths")
fig = go.Figure(data=[go.Table(
    header=dict(values=[['<b>STATE</b>'],['<b>INFLUENZA DEATHS</b>']],
                fill_color='seagreen',
                font=dict(color='white'),
                align='center'),
    cells=dict(values=[df_3.state, df_3['influenza_deaths']],
               fill_color='lightgreen',
               align='center'))
])

fig.show()
```

STATE	INFLUENZA DEATHS
New York Total	1081
California Total	555
Texas Total	315
Florida Total	289
Ohio Total	236

Covid19 comparison

In the plot below, we are doing a side-by-side comparison of the total number of cases, total number of deaths and number of people who have recovered from Covid19 in all the states of US.

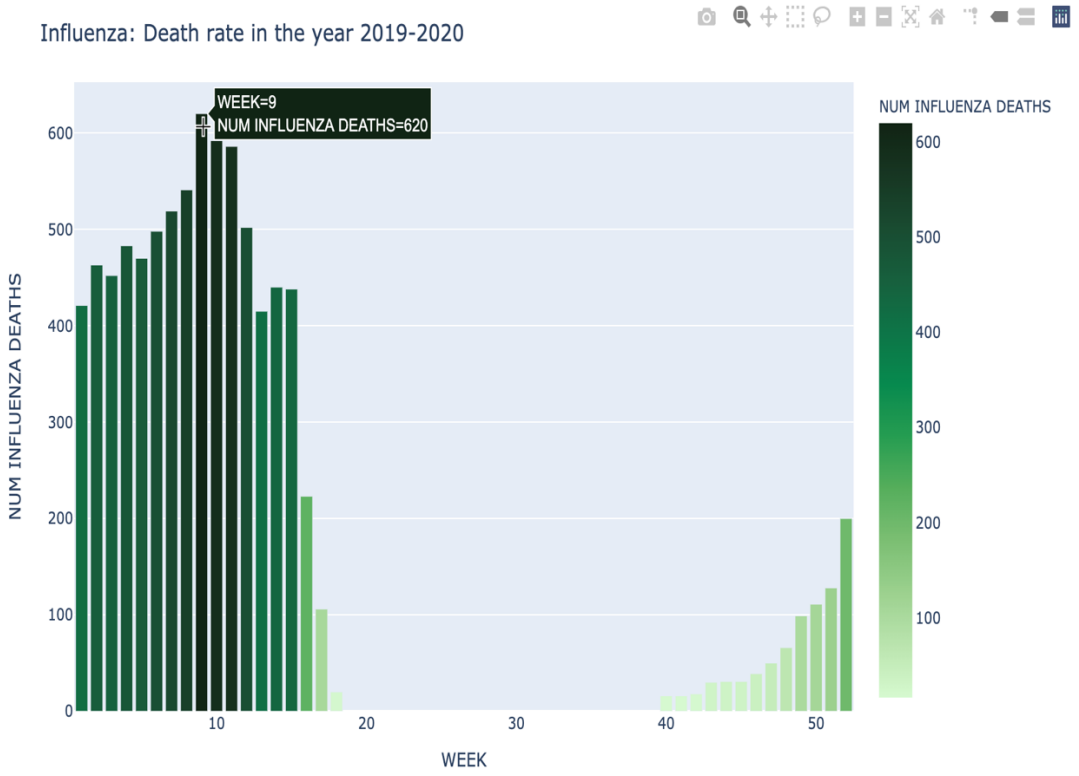
```
In [99]: #import plotly.graph_objects as go

         fig = go.Figure()
         fig.add_trace(go.Bar(
             y=df_usa_covid19['Province_State'],
             x=df_usa_covid19['Confirmed'],
             name='Covid19 Confirmed',
             orientation='h',
             marker=dict(
                 color='rgb(255,255,0)',
             )
         ))
         fig.add_trace(go.Bar(
             y=df_usa_covid19['Province_State'],
             x=df_usa_covid19['Recovered'],
             name='Covid19 Recovered',
             orientation='h',
             marker=dict(
                 color='rgb(0,255,0)',
             )
         ))
         fig.add_trace(go.Bar(
             y=df_usa_covid19['Province_State'],
             x=df_usa_covid19['Deaths'],
             name='Covid19 Deaths',
             orientation='h',
             marker=dict(
                 color='rgb(128,0,0)',
             )
         ))
         fig.update_layout(title_text = 'COVID19: Total cases, Recovered and Deaths from Feb\'20-May\'20',height=1000,barmode='s
         fig.show()
```



Bar plot of the Influenza deaths in US over the year 2019

The Bar plot below represent the intensity of weekly increase and decrease of Influenza deaths in the year 2019 in US. So, from the graph below we can infer that influenza deaths are more prevalent during the 9^{th}-11^{th} week of a year.

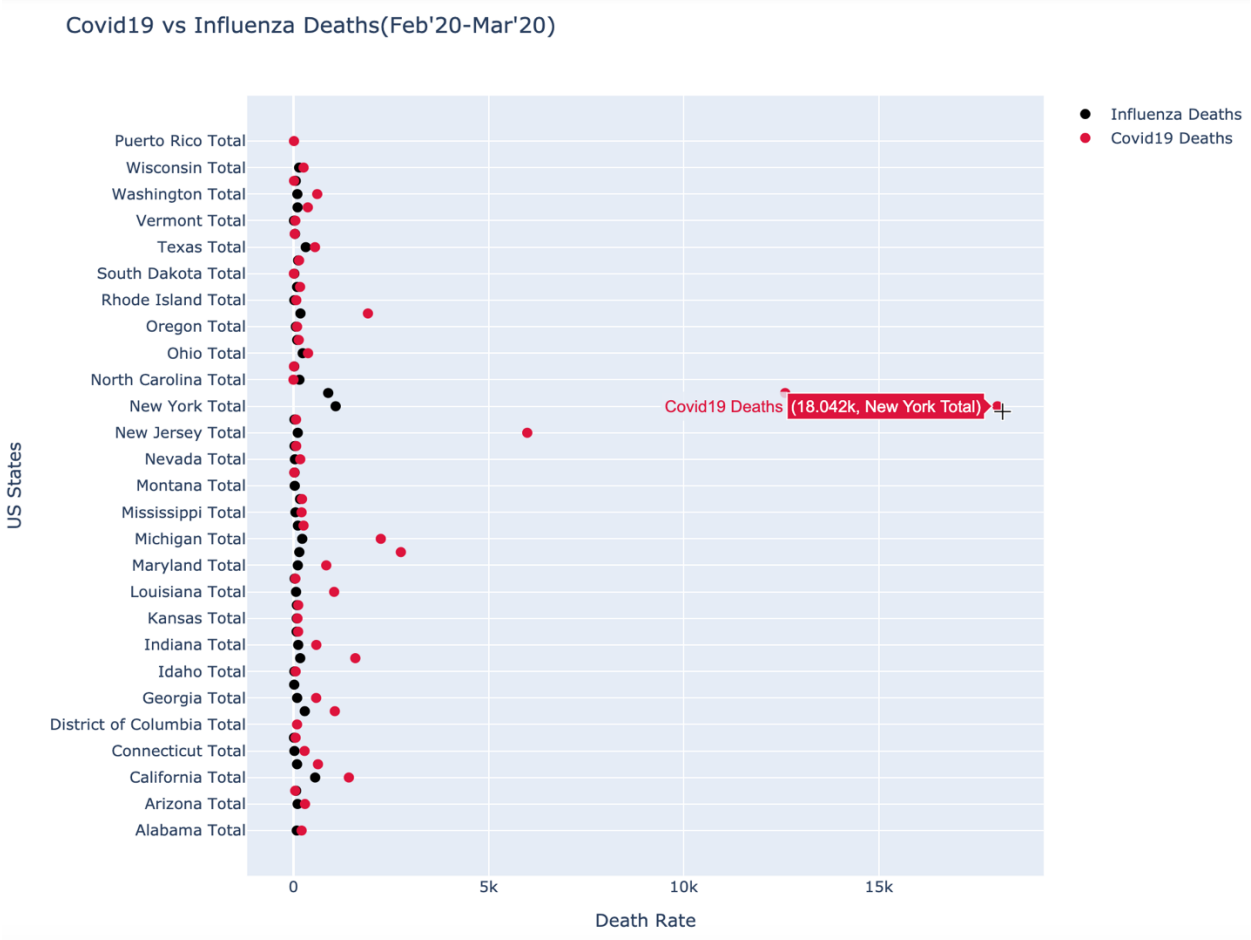
```
In [21]: data_usa_flu = pd.read_csv('National_Custom_Data.csv')
         fig = px.bar(data_usa_flu, x='WEEK', y='NUM INFLUENZA DEATHS',
                      hover_data=['WEEK', 'NUM INFLUENZA DEATHS'], color='NUM INFLUENZA DEATHS',
                      labels={'Influenza: Death rate in the year 2019-2020'}, height=600, color_continuous_scale='algae')
         fig.update_layout(title_text = 'Influenza: Death rate in the year 2019-2020')
         fig.show()
```



Scatterplot comparison

Here we are comparing the Influenza death intensity with Covid19 from Feb'20-May'20. The black dots represent the Influenza deaths whereas the crimson color represents the Covid19 deaths. We can clearly interpret from the plot that Covid19 has been deadlier than flu deaths. Covid19 cases and deaths have risen exponentially over the past 3months.

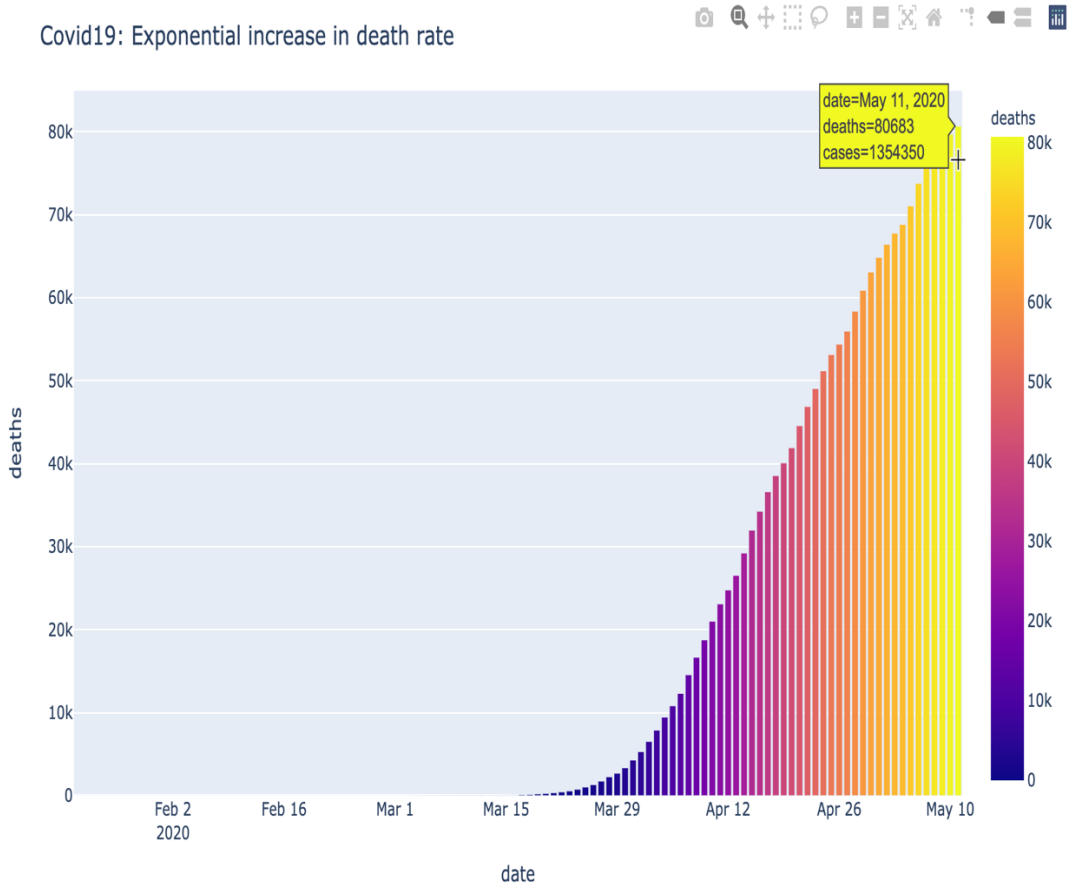
In [139]:
```
fig = go.Figure()
fig.add_trace(go.Scatter(
    x=dfFinal_Covid_total['influenza_deaths'],
    y=dfFinal_Covid_total['state'],
    marker=dict(color="black", size=8),
    mode="markers",
    name="Influenza Deaths",
))

fig.add_trace(go.Scatter(
    x=dfFinal_Covid_total['covid-19_deaths'],
    y=dfFinal_Covid_total['state'],
    marker=dict(color="crimson", size=8),
    mode="markers",
    name="Covid19 Deaths",
))

fig.update_layout(title="Covid19 vs Influenza Deaths(Feb'20-Mar'20)",
                  xaxis_title="Death Rate",
                  yaxis_title="US States",
                  height=800)

fig.show()
```



Exponential increase in death rate: Covid19

The below plot shows an explosive growth of Coronavirus deaths in US from Feb'20-May'20. As the hover data shows us, the total death count has crossed the 80000 mark till the present day in just 3.5months.

```
In [20]: data_usa = pd.read_csv('https://raw.githubusercontent.com/nytimes/covid-19-data/master/us.csv')
fig = px.bar(data_usa, x='date', y='deaths',

             hover_data=['cases', 'deaths'], color='deaths',
             labels={'Covid19: Exponential increase in death rate'}, height=600)
fig.update_layout(title_text = 'Covid19: Exponential increase in death rate')
fig.show()
```



Part 3(Hypothesis Testing)

In the third part of the project, we will be performing Hypothesis testing to determine if the null hypothesis should be rejected because its probability is below a predetermined significance level.

Test Conditions

Null Hypothesis (H_0) : Covid19 is not any more dangerous than the normal flu (We will compare the death counts) Covid19 = Influenza

Alternate Hypothesis (H_a) : Covid19 is more dangerous that the common influenza

Significance Level Alpha: 0.05

Reject the Null Hypothesis if Probability > .05 (Significance Level)

We will leverage 2 Sample t-tests to test our Hypothesis:

A Two-Sample t-test is used to test the difference between two population means. A common application is to determine whether the 2 means are equal. In our application of the Two-Sample t-test, we will be taking a 3 States (1, NY, 2-Illonis, 3-Washington) which is a good representation of the geographical disparity in the United States. For each state, we will compare the means of deaths relegated to Covid19 and death related to Influenza and calculate the probability of each mean being equal. If the probability is less than the significance level alpha = .05, we will reject the Null Hypothesis.

Imports & preprocessing the dataset

Library used for Hypothesis Testing:

- Pandas

Then we read the dataset with the weekly Covid19 and Influenza deaths across the US states in the from Feb'20-May'20 and placed it in a dataframe.

In [2]:
```
import pandas as pd

deaths_df = pd.ExcelFile('D:\Downloads\weekly state-wise death by covid19 and flu 2020.xls')
deaths_df = pd.read_excel(deaths_df,0,skiprows = 0)
deaths_df['State'].unique()

# deaths_df.fillna(0)
```

Out[2]:
```
array(['United States', 'Alabama', 'Alaska', 'Arizona', 'Arkansas',
       'California', 'Colorado', 'Connecticut', 'Delaware',
       'District of Columbia', 'Florida', 'Georgia', 'Hawaii', 'Idaho',
       'Illinois', 'Indiana', 'Iowa', 'Kansas', 'Kentucky', 'Louisiana',
       'Maine', 'Maryland', 'Massachusetts', 'Michigan', 'Minnesota',
       'Mississippi', 'Missouri', 'Montana', 'Nebraska', 'Nevada',
       'New Hampshire', 'New Jersey', 'New Mexico', 'New York',
       'New York City', 'North Carolina', 'North Dakota', 'Ohio',
       'Oklahoma', 'Oregon', 'Pennsylvania', 'Rhode Island',
       'South Carolina', 'South Dakota', 'Tennessee', 'Texas', 'Utah',
       'Vermont', 'Virginia', 'Washington', 'West Virginia', 'Wisconsin',
       'Wyoming', 'Puerto Rico'], dtype=object)
```

Plotting the Distribution of Covid19 and Influenza

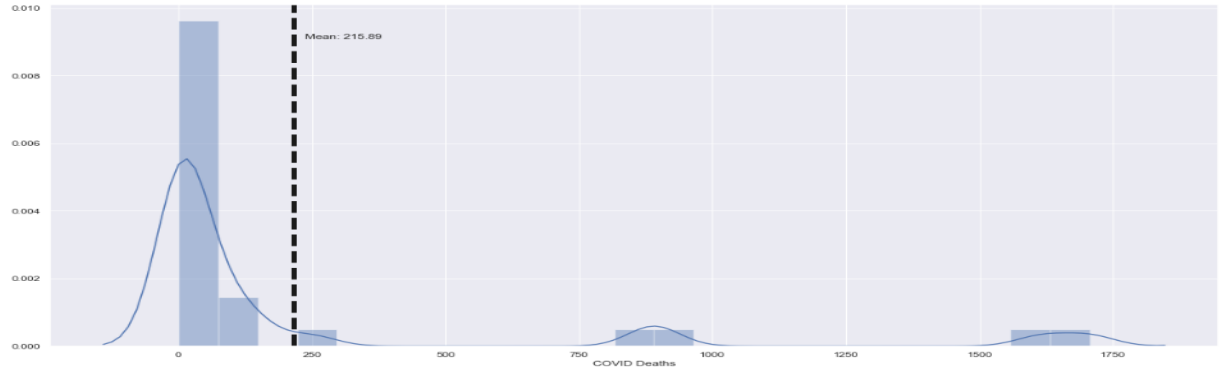
Libraries used:

- Matplotlib
- Seaborn

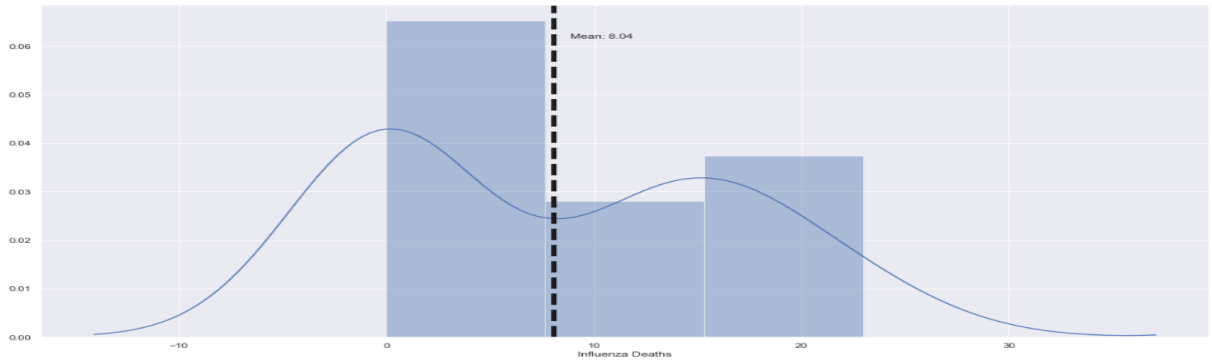
```
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
from pylab import rcParams
%matplotlib inline
warnings.filterwarnings("ignore")
rcParams['figure.figsize'] = 20,10
rcParams['font.size'] = 30
sns.set()

def plot_distribution(inp):
    plt.figure()
    ax = sns.distplot(inp)
    plt.axvline(np.mean(inp), color="k", linestyle="dashed", linewidth=5)
    _, max_ = plt.ylim()
    plt.text(
        inp.mean() + inp.mean() / 10,
        max_ - max_ / 10,
        "Mean: {:.2f}".format(inp.mean()),
    )
    return plt.figure

plot_distribution(covid_deaths)
plot_distribution(influenza_deaths)
```

Distribution plots of Sample 1 T-Test (New York)

New York Covid-19 Distribution of Deaths



New York Influenza Distribution of Deaths



Sample 1 T-Test (New York)

```python
from scipy.stats import f_oneway
from scipy.stats import ttest_ind

def compare_2_groups(arr_1, arr_2, alpha, sample_size):
    stat, p = ttest_ind(arr_1, arr_2)
    print('Statistics=%.3f, p=%.3f' % (stat, p))
    if p > alpha:
        print('Same distributions (fail to reject H0)')
    else:
        print('Different distributions (reject H0)')


sample_size = 14
covid_deaths= newyork_covid_deaths['COVID Deaths']
influenza_deaths = newyork_influenza_deaths['Influenza Deaths']
compare_2_groups(covid_deaths, influenza_deaths, 0.05, sample_size)

```

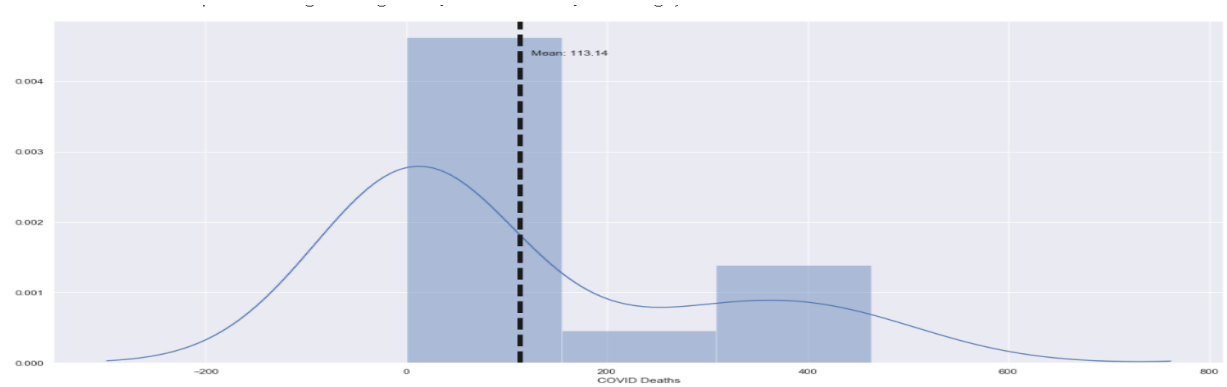
```
Statistics=2.253, p=0.033
Different distributions (reject H0)
```

Result: p value is 0.033 < 0.05(significance level)

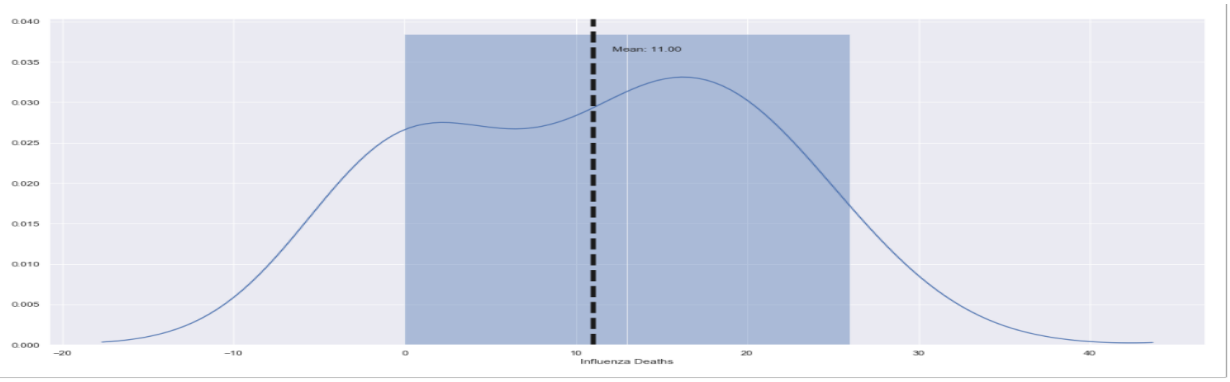
Reject Null Hypothesis for Sample 1-New York

Distribution plots of Sample 2 T-Test (Illinois)

Illinois Covid-19 Distribution of Deaths



Illinois Influenza- Distribution of Deaths



Sample 2 T-Test (Illinois)

In [233]:

```
from scipy.stats import f_oneway
from scipy.stats import ttest_ind

def compare_2_groups(arr_1, arr_2, alpha, sample_size):
    stat, p = ttest_ind(arr_1, arr_2)
    print('Statistics=%.3f, p=%.3f' % (stat, p))
    if p > alpha:
        print('Same distributions (fail to reject H0)')
    else:
        print('Different distributions (reject H0)')


sample_size = 14
covid_deaths= illinois_covid_deaths['COVID Deaths']
influenza_deaths = illinois_influenza_deaths['Influenza Deaths']
compare_2_groups(covid_deaths, influenza_deaths, 0.05, sample_size)

```

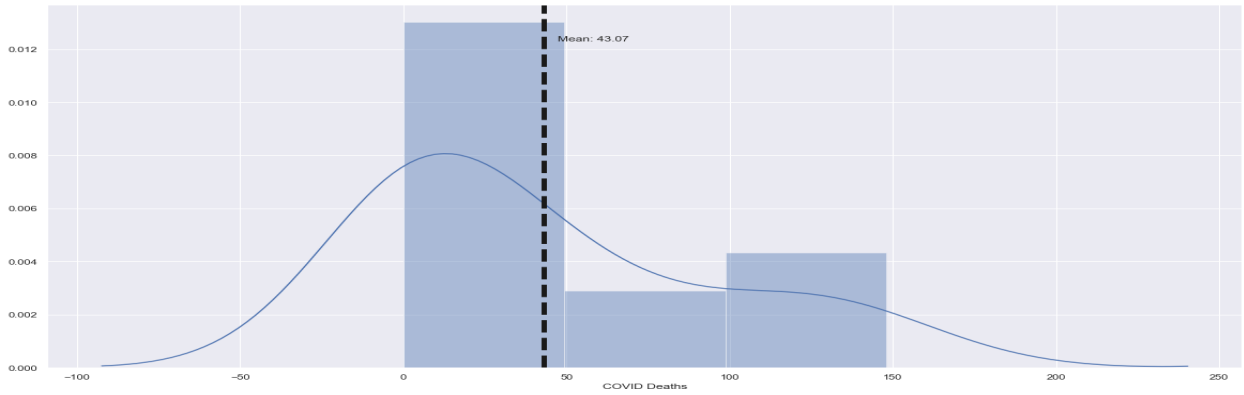
```
Statistics=2.216, p=0.036
Different distributions (reject H0)
```

Result: P-Value of 2 Sample T-test of Illinois is 0.036 < 0.05(significance level)

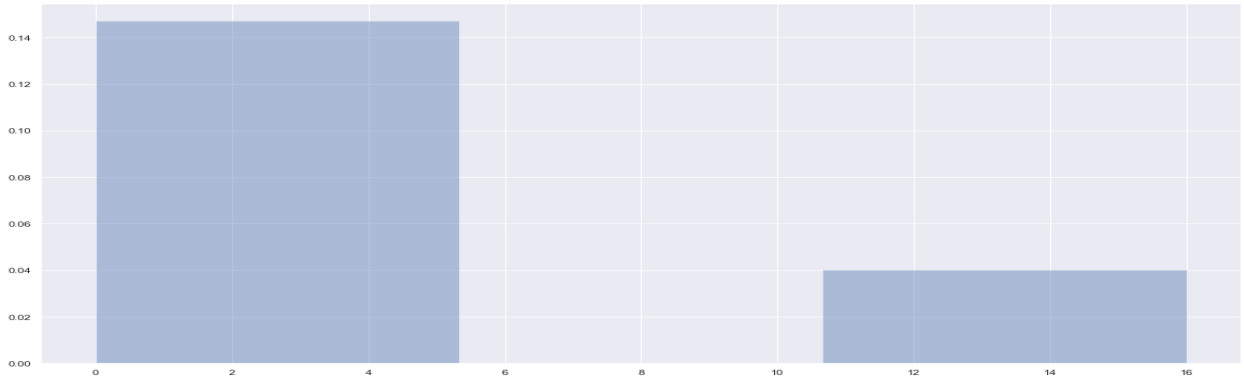
Reject Null Hypothesis for Sample 2-Illinois

Distribution plots of Sample 3 T-Test (Washington)

Washington State Covid-19 Distribution of Deaths



Washington State- Influenza Distribution of Deaths



Sample 3 T-Test (Washington)

In [253]:

```
from scipy.stats import f_oneway
from scipy.stats import ttest_ind

def compare_2_groups(arr_1, arr_2, alpha, sample_size):
    stat, p = ttest_ind(arr_1, arr_2)
    print('Statistics=%.3f, p=%.3f' % (stat, p))
    if p > alpha:
        print('Same distributions (fail to reject H0)')
    else:
        print('Different distributions (reject H0)')


sample_size = 14
covid_deaths= state_covid_deaths['COVID Deaths']
influenza_deaths = state_influenza_deaths['Influenza Deaths']
compare_2_groups(covid_deaths, influenza_deaths, 0.05, sample_size)

```

```
Statistics=2.878, p=0.008
Different distributions (reject H0)
```

Result: P=Value of Sample 3 T-Test of Washington is 0.008< 0.05(Significance level)

Reject Null Hypothesis for Sample 3-Washington State

Hypothesis Testing Inference

In all the 3 samples, there was supporting evidence to reject the null hypothesis and supporting the alternate hypothesis that Covid19 is not like the common flu and that it causes a larger number of deaths.

CONCLUSION

In this project, we deep-dived and performed an exhaustive analysis in support of our goal in finding supportive evidence to reject the notion that the Covid-19 Pandemic is just like Influenza (the common flu). We gathered data for dispersed sources in support of our EDA (Exploratory Data Analysis) and subsequently and most importantly the Hypothesis Testing to reject null Hypothesis. The highlights of the EDA included a 5-point summary which showed that the average rate of deaths for Covid-19 (1067/day) was substantially higher than the flu (129/day at peak season). Furthermore, the heat plots for both Covid-19 and Influenza depicted that both of these illnesses are quite well spread throughout the USA but with consistency in high intensity with the epicenter of NY in both cases. Additional heat maps for both illnesses indicated a much higher rate of death of Covid-19. PyPlot's stacked bar chart provided us a state-wide view for Confirmed Cases, Death and Recovery. Further providing support for the wide-spread nature of Covid-19. We concluded our EDA analysis with a Top N analysis chart with top 5 states for number of deaths for each category. The Top 5 chart painted a daunting picture of Covid-19 deaths in comparison to the flu.

The crux of our supporting evidence to reject our null hypothesis that Covid-19 is just like any other flu was the results of Hypothesis testing using a 2 Sample t-test. What could be worse than death? So, we compared 3 Samples of deaths due to Covid-19 and Influenza (New York, Illinois and Washington State) representing each region of the country to prevent any biasing of data because of regional demographic anomalies to highlight the severity of the Covid-19 impact. The results of the Hypothesis testing were overwhelming in favor of rejecting our null Hypothesis and in support of Alternate Hypothesis that Coivd19 is much more dangerous in terms of number of deaths. In all 3 samples the probability was less than the Significant value of 0.05 in rejecting the null Hypothesis.

In closing, Covid-19 is a serious illness that is causing 10 times more death than Influenza in a short period of time. It would be ill advised to compare this to the common flu as demonstrated herein in the project.