Project - COVID-19 New Jersey Trends & Impact on RideSharing Platform

In [1]:
```
# Mount your google drive where you've saved your assignment folder
# from google.colab import drive
# drive.mount('/content/gdrive')
```

In [2]:
```
# cd '/content/gdrive/My Drive/CSE544_project_112669645/'
```

In [3]:
```
# pip install dexplot
```

In [4]:
```
import pandas as pd
# import the seaborn module
import seaborn as sns
import matplotlib.pyplot as plt
import datetime  as dt
import numpy as np
from matplotlib.ticker import PercentFormatter
import os
import missingno as msno # visualize the distribution of NaN values
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline
import plotly
from datetime import datetime
import dexplot as dxp
import plotly.graph_objects as go
from plotly.subplots import make_subplots
import plotly.express as px
```

COVID-19 Dataset --> We have taken New Jersey covid19 data source --> https://covidtracking.com/api/v1/states/daily.csv (https://covidtracking.com/api/v1/states/daily.csv)

X Dataset --> We are trying to observe the impact of COVID-19 on the stock prices of major Ridesharing Players (Uber + Lyft)

https://finance.yahoo.com/quote/UBER/history?p=UBER (https://finance.yahoo.com/quote/UBER/history?p=UBER)

https://finance.yahoo.com/quote/LYFT/history?p=LYFT (https://finance.yahoo.com/quote/LYFT/history?p=LYFT)

Project Git Repository --> https://github.com/marif1901/COVID19_NJ_ImpactAnalysis (https://github.com/marif1901/COVID19_NJ_ImpactAnalysis)

Part 1: Data Pre Processing (10%)

In [5]:
```
cov_url= 'https://raw.githubusercontent.com/marif1901/COVID19_NJ_ImpactAnalysis/master/COVID19_NJ_Data.csv'
x_uber_url= "https://raw.githubusercontent.com/marif1901/COVID19_NJ_ImpactAnalysis/master/UBER_1Y.csv"
x_lyft_url= "https://raw.githubusercontent.com/marif1901/COVID19_NJ_ImpactAnalysis/master/LYFT_1Y.csv"
```

Reading Datasets

In [6]:
```
covid = pd.read_csv(cov_url,sep=',')# use sep="," for coma separation.
xuber = pd.read_csv(x_uber_url,sep=',')
xlyft = pd.read_csv(x_lyft_url,sep=',')
print(covid.columns)
print(xuber.columns)
print(xlyft.columns)
```

```
Index(['date', 'state', 'positive', 'negative', 'pending',
       'hospitalizedCurrently', 'hospitalizedCumulative', 'inIcuCurrently',
       'inIcuCumulative', 'onVentilatorCurrently', 'onVentilatorCumulative',
       'recovered', 'dataQualityGrade', 'lastUpdateEt', 'hash', 'dateChecked',
       'death', 'hospitalized', 'total', 'totalTestResults', 'posNeg', 'fips',
       'deathIncrease', 'hospitalizedIncrease', 'negativeIncrease',
       'positiveIncrease', 'totalTestResultsIncrease', 'dailypositvecases',
       'dailynegativecases', 'dailytestingdone', 'dailydeath'],
      dtype='object')
Index(['Date', 'Open', 'High', 'Low', 'Close', 'Adj Close', 'Volume'], dtype='object')
Index(['Date', 'Open', 'High', 'Low', 'Close', 'Adj Close', 'Volume'], dtype='object')
```

Preprocessing on COVID Data

In [7]:
```
covid_cols= ['date','dailypositvecases','dailynegativecases','dailydeath','dailytestingdone',
             'positiveIncrease','negativeIncrease', 'deathIncrease','totalTestResultsIncrease',
            'positive', 'negative', 'death','totalTestResults']
covid_sel= covid[covid_cols].copy()

covid_cols= ['date','dailypositvecases','dailynegativecases','dailydeath','dailytestingdone',
             'positiveIncrease','negativeIncrease', 'deathIncrease','totalTestResultsIncrease',
            'cumpositive', 'cumnegative', 'cumdeath','cumtotalTestResults']

covid_sel.columns=  covid_cols
```

Dropping rows where data is NA

In [8]:
```
count_nulls= sum(pd.isna(covid_sel['date']))
print('\033[1m' + ' Total nulls found :' + str(count_nulls))
index = covid_sel[pd.isna(covid_sel['date'])].index
covid_sel.drop(index , inplace=True)
```

```
 Total nulls found :0
```

Converting date to proper %Y%m%d format

In [9]:
```
covid_sel['date']= covid_sel['date'].astype(str)
covid_sel['date'] = pd.to_datetime(covid_sel['date'], format='%Y%m%d').dt.strftime("%Y-%m-%d");
```

In [10]:
```
int_col= ['dailypositvecases','dailynegativecases','dailydeath','dailytestingdone',
             'positiveIncrease','negativeIncrease', 'deathIncrease','totalTestResultsIncrease',
            'cumpositive', 'cumnegative', 'cumdeath','cumtotalTestResults']
covid_sel[int_col] = covid_sel[int_col].astype(np.int32)
covid_sel.head(3)
```

Out[10]:

	date	dailypositvecases	dailynegativecases	dailydeath	dailytestingdone	positiveIncrease	negativeIncrease	deathIncrease	totalTestResultsIncrease	cumpositive	cumnegative	cumdeath	cumtotalTestResults
0	2020-05-07	68760	90580	4341	159340	1745	1993	252	3738	133635	159023	8801	292658
1	2020-05-06	64875	68443	4460	133318	1297	0	305	1297	131890	157030	8549	288920
2	2020-05-05	67015	88587	4089	155602	2324	8079	334	10403	130593	157030	8244	287623

In [11]:
```
print('\033[1m' +'Min Date observed for COVID : ' + str(covid_sel['date'].min()))
print('\033[1m' + 'Max Date observed for COVID: ' + str(covid_sel['date'].max()))
```

Min Date observed for COVID : 2020-03-05
Max Date observed for COVID: 2020-05-07

Preprocessing on X Data

In [12]:
```
x_cols= ['Date','Close','Volume']

xuber_sel= xuber[x_cols].copy()
xlyft_sel= xlyft[x_cols].copy()

x_cols= ['date','UberClosingPrice','UberVolume']
xuber_sel.columns= x_cols

x_cols= ['date','LyftClosingPrice','LyftVolume']
xlyft_sel.columns=x_cols
```

In [13]:
```
xuber_sel.date= pd.to_datetime(xuber_sel['date']).dt.strftime('%Y-%m-%d')
xlyft_sel.date=pd.to_datetime(xlyft_sel['date']).dt.strftime('%Y-%m-%d')
```

In [14]:
```
x_sel= pd.merge(xuber_sel, xlyft_sel,on='date')
print('\033[1m' + 'Min Date observed for X : ' + str(x_sel['date'].min()))
print('\033[1m' + 'Max Date observed for X: ' + str(x_sel['date'].max()))
```

Min Date observed for X : 2019-05-10
Max Date observed for X: 2020-05-07

In [15]:
```
x_sel.head(3)
```

Out[15]:

	date	UberClosingPrice	UberVolume	LyftClosingPrice	LyftVolume
0	2019-05-10	41.570000	186322500	51.090000	23111200
1	2019-05-13	37.099998	79442400	48.150002	10007400
2	2019-05-14	39.959999	46661100	50.520000	7007400

Merging COVID data with X Data for Analysing impact in the same time frame

In [16]:
```
comb_df= covid_sel.merge(x_sel, how='inner', on='date')
comb_df=comb_df.drop_duplicates()
print('\033[1m' + 'Min Date observed for comb_df : ' + str(comb_df['date'].min()))
print('\033[1m' + 'Max Date observed for comb_df: ' + str(comb_df['date'].max()))
```

```
Min Date observed for comb_df : 2020-03-05
Max Date observed for comb_df: 2020-05-07
```

Filtering 8 weeks timeframe for Analysis, Starting Date from. Monday 9th March, End Date Sunday 3rd May

In [17]:
```
st_dt= pd.to_datetime('2020-03-09').strftime('%Y-%m-%d')
# print(st_dt)
end_dt= pd.to_datetime('2020-05-04').strftime('%Y-%m-%d')
# print(end_dt)

comb_df = comb_df[ (comb_df['date']>=st_dt) & (comb_df['date']<= end_dt)]

print('\033[1m' + 'Min Date observed for comb_df : ' + str(comb_df['date'].min()))
print('\033[1m' + 'Max Date observed for comb_df: ' + str(comb_df['date'].max()))
print('\033[1m' + 'Total Rows * cols: ' + str(comb_df.shape))

comb_df.head(3)
```

```
Min Date observed for comb_df : 2020-03-09
Max Date observed for comb_df: 2020-05-04
Total Rows * cols: (40, 17)
```

Out[17]:

	date	dailypositvecases	dailynegativecases	dailydeath	dailytestingdone	positiveIncrease	negativeIncrease	deathIncrease	totalTestResultsIncrease	cumpositive	cumnegative	cumdeath	cumtotalTestResults	UberClosin
3	2020-05-04	63578	68443	4155	132021	1525	629	39	2154	128269	148951	7910	277220	27.4
4	2020-05-01	61664	70781	3626	132445	2538	6089	310	8627	121190	135355	7538	256545	28.3
5	2020-04-30	59526	64574	3912	124100	2388	4212	458	6600	118652	129266	7228	247918	30.2

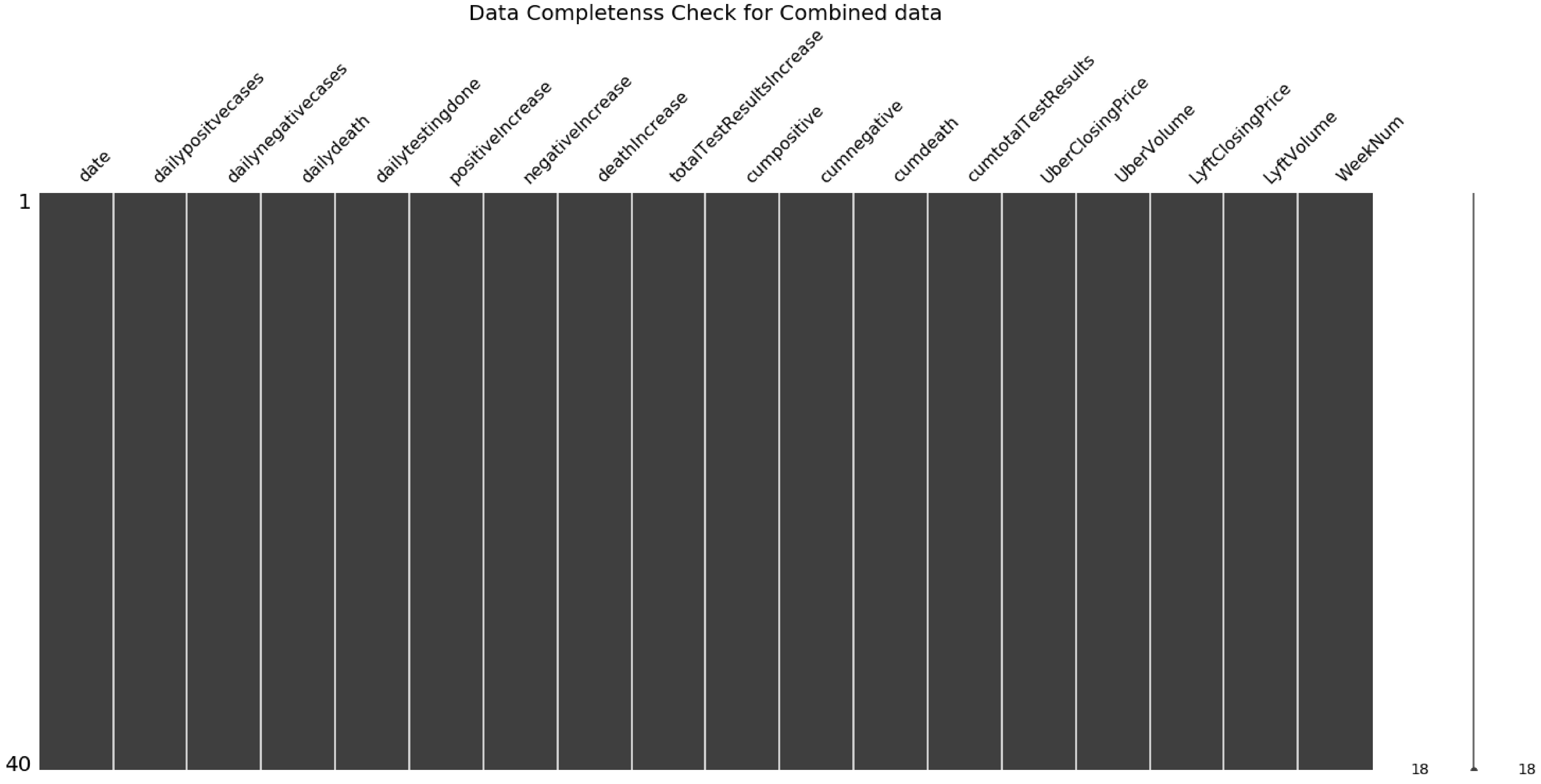
Assigning Week Number

In [18]:
```
comb_df['WeekNum'] = ((pd.to_datetime(comb_df['date']) - pd.to_datetime(st_dt)).dt.days)//7 +1
```

Checking Nullity and Data Completeness

In [19]:
```
msno.matrix(comb_df)
plt.title('Data Completenss Check for Combined data', size = 20)
```

Out[19]: Text(0.5, 1.0, 'Data Completenss Check for Combined data')



No Nullity found above

Let's Apply the Tukey's Rule to check if there are any data Outliers

In [20]:
```
Q1 = comb_df.quantile(0.25)
Q3 = comb_df.quantile(0.75)
IQR = Q3 - Q1
print(IQR.astype(np.int32))
print('\033[1m' + 'shape before Outlier Detection' + str(comb_df.shape))
```

```
dailypositvecases              44245
dailynegativecases             44204
dailydeath                      2337
dailytestingdone               88831
positiveIncrease                2746
negativeIncrease                3503
deathIncrease                    300
totalTestResultsIncrease        6036
cumpositive                    87345
cumnegative                    89712
cumdeath                        4448
cumtotalTestResults           177058
UberClosingPrice                   3
UberVolume                  17006075
LyftClosingPrice                   6
LyftVolume                   6008325
WeekNum                            4
dtype: int32
shape before Outlier Detection(40, 18)
```

In [21]:
```
comb_out = comb_df[~((comb_df < (Q1 - 1.5 * IQR)) |(comb_df > (Q3 + 1.5 * IQR))).any(axis=1)]
print('\033[1m' + 'shape after Outlier Detection' + str(comb_out.shape))
# comb_df= comb_out.copy()
```

```
shape after Outlier Detection(36, 18)
```

We can see that after Outlier detectin we are left with 36 rows, 4 rows are deleted

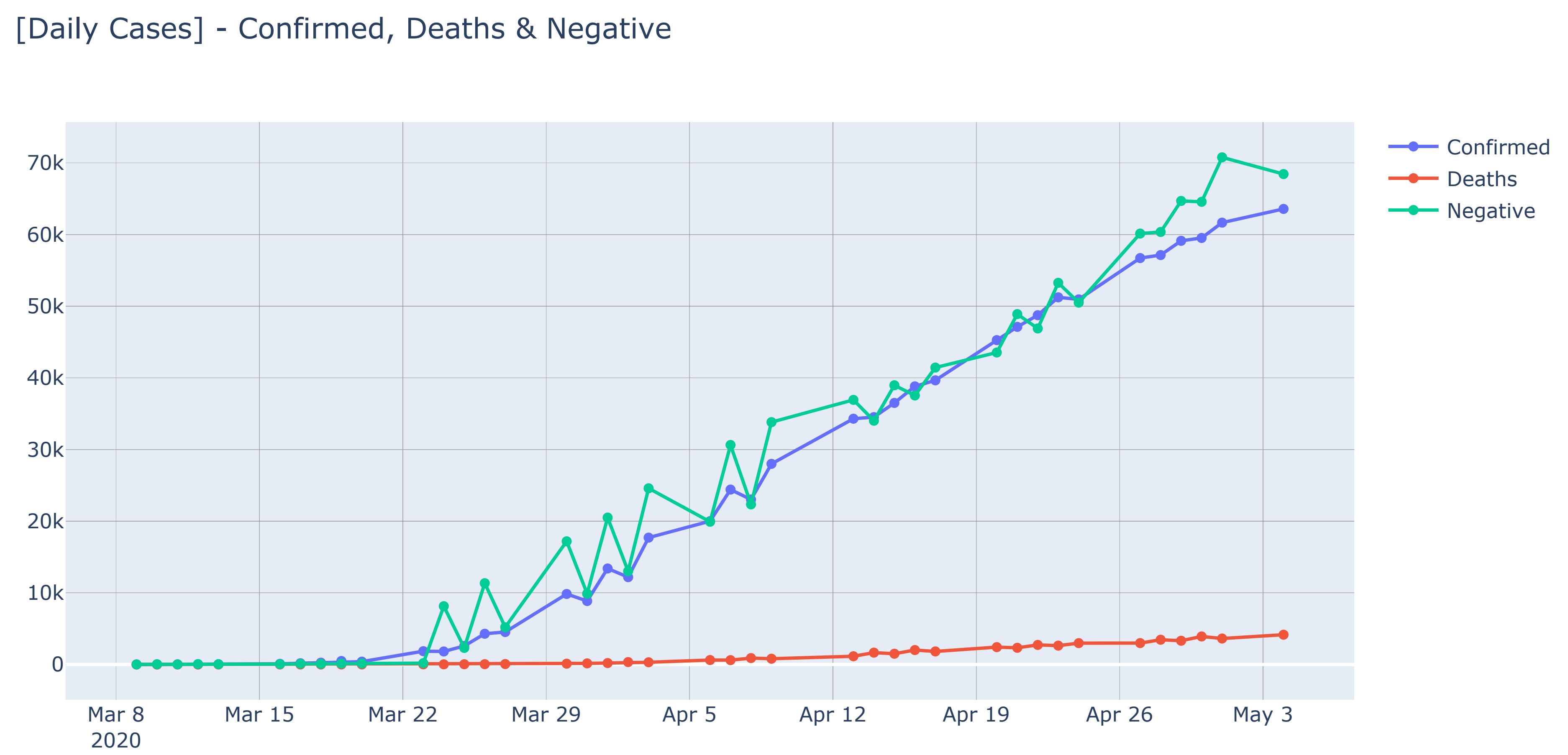
In [22]:
```
comb_df= comb_df.sort_values(by="date")
print(comb_df.shape)
```

```
(40, 18)
```

Part 2: General Trends in Covid + X Data (10%)

Day on Day Trends | PDF | CDF of COVID 19 Growth

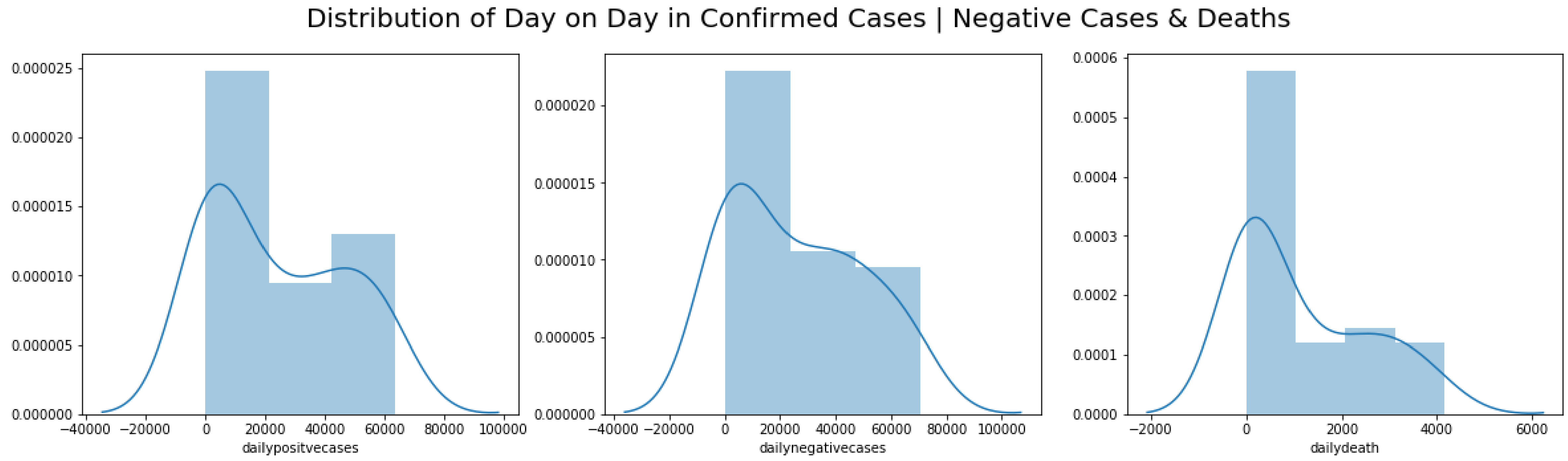
In [23]:
```
fig = go.Figure()
fig.add_trace(go.Scatter(x=comb_df['date'], y=comb_df['dailypositivecases'],
                         mode='lines+markers', name='Confirmed'))
fig.add_trace(go.Scatter(x=comb_df['date'], y=comb_df['dailydeath'],
                         mode='lines+markers', name='Deaths'))
fig.add_trace(go.Scatter(x=comb_df['date'], y=comb_df['dailynegativecases'],
                         mode='lines+markers', name='Negative'))

fig.update_layout(
        xaxis_title="",
        yaxis_title="",
        title = '[Daily Cases] - Confirmed, Deaths & Negative'
#         yaxis_type="log"
    )
fig.show()
```



Let's check the distribution of data for Confirmed Cases, Negative Cases and Deaths

In [24]:
```
#histogram
fig = plt.figure(figsize= (20,5))
plt.subplot(1,3,1)
sns.distplot(comb_df['dailypositvecases'])

plt.subplot(1,3,2)
sns.distplot(comb_df['dailynegativecases'])

plt.subplot(1,3,3)
sns.distplot(comb_df['dailydeath'])

fig.suptitle("Distribution of Day on Day in Confirmed Cases | Negative Cases & Deaths", fontsize=20)
```

Out[24]:
```
Text(0.5, 0.98, 'Distribution of Day on Day in Confirmed Cases | Negative Cases & Deaths')
```

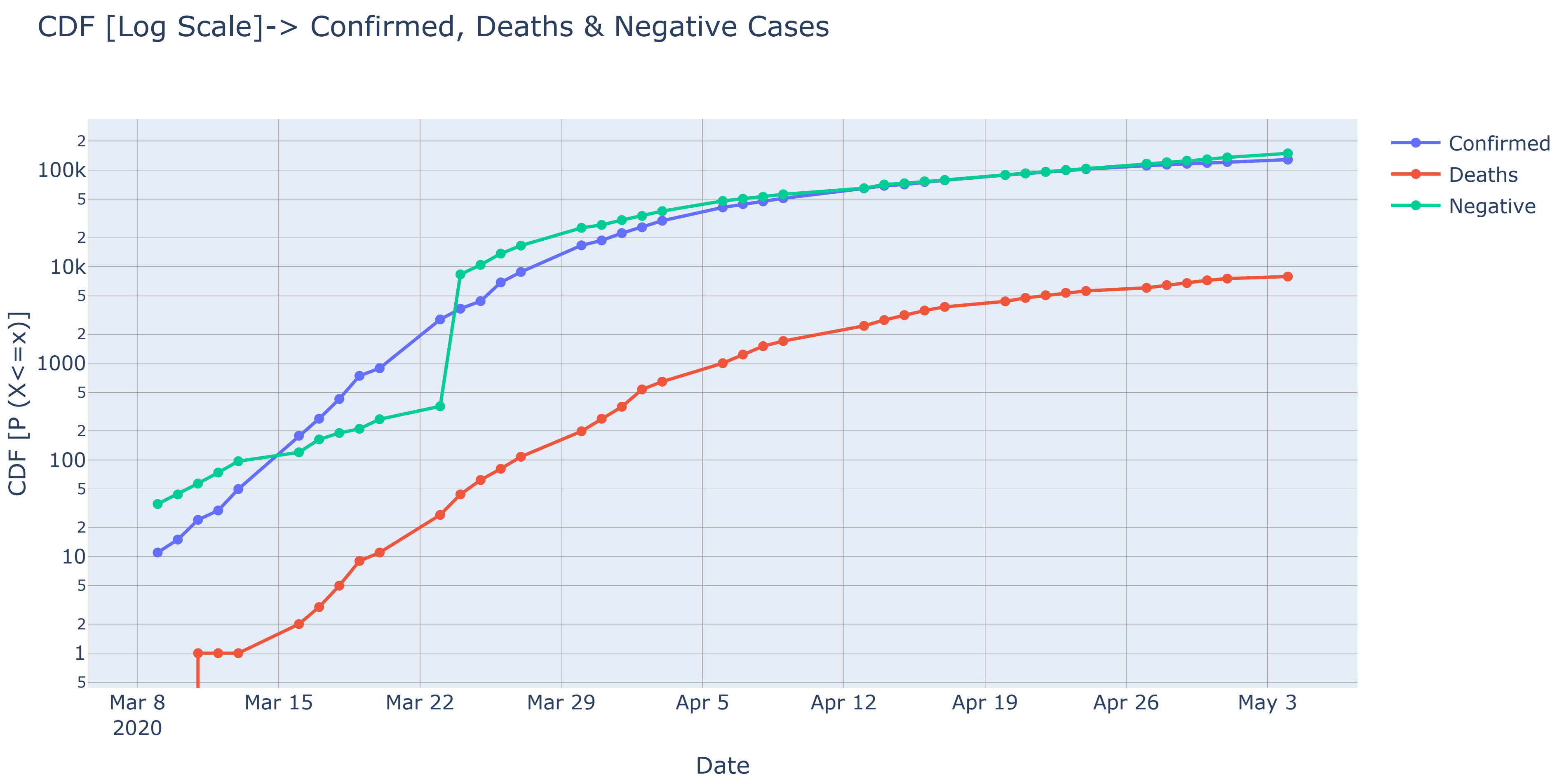


Inference from above graph: we can clearly see that for confirmed and negative cases it follows a smooth curve with fluctuations while death is mostly uniform after certain number of days so its flat in nature

---

"CURVE IS FLATTENING" after 2 Months ??

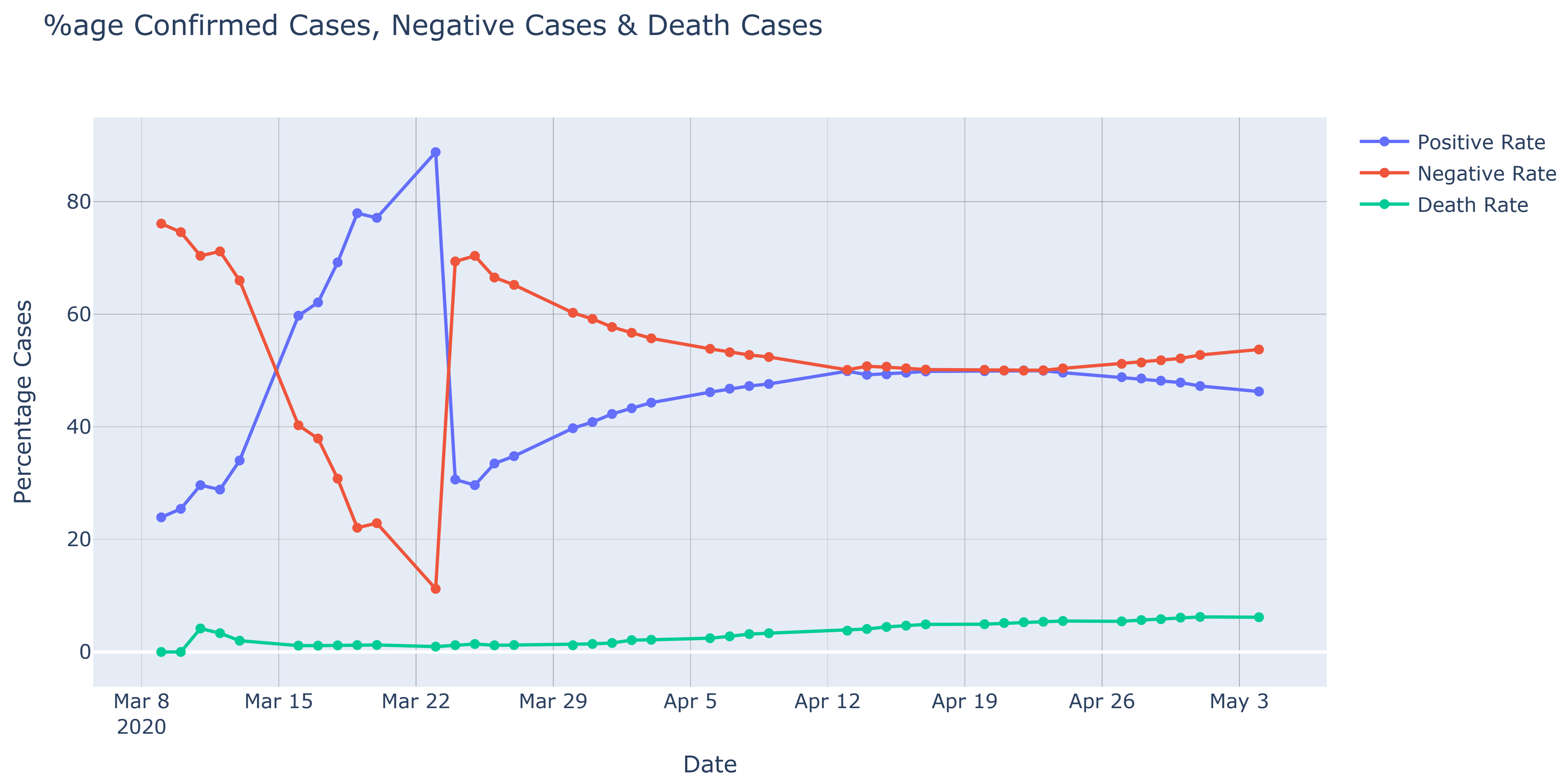
In [25]:
```
fig = go.Figure()
fig.add_trace(go.Scatter(x=comb_df['date'], y=comb_df['cumpositive'],
                         mode='lines+markers', name='Confirmed'))
fig.add_trace(go.Scatter(x=comb_df['date'], y=comb_df['cumdeath'],
                         mode='lines+markers', name='Deaths'))
fig.add_trace(go.Scatter(x=comb_df['date'], y=comb_df['cumnegative'],
                         mode='lines+markers', name='Negative'))

fig.update_layout(
        xaxis_title="Date",
        yaxis_title="CDF [P (X<=x)]",
#         title = 'Cumulative -> Confirmed, Deaths & Negative Results'
        title = 'CDF [Log Scale]-> Confirmed, Deaths & Negative Cases',
        yaxis_type="log"
    )
fig.show()
```



Inference from above graph: It can be observed there was a steep increase in the confirm cases from Mar9 to Apr6 since then the rate of increase seems to be decreasing and curve looks to be flattening after Apr20 while death is observed to be increasing at constant pace

What are the Percentage Mix of Postive | Negative | Death Cases ??

In [26]:
```
df_t= comb_df.copy()
df_t['Positive Rate'] = df_t['cumpositive']*100/df_t['cumtotalTestResults']
df_t['Negative Rate'] = df_t['cumnegative']*100/df_t['cumtotalTestResults']
df_t['Death Rate'] = df_t['cumdeath']*100/df_t['cumpositive']
fig = go.Figure()
fig.add_trace(go.Scatter(x=df_t['date'], y=df_t['Positive Rate'], mode='lines+markers', name='Positive Rate'))
fig.add_trace(go.Scatter(x=df_t['date'], y=df_t['Negative Rate'], mode='lines+markers', name='Negative Rate'))
fig.add_trace(go.Scatter(x=df_t['date'], y=df_t['Death Rate'], mode='lines+markers', name='Death Rate'))
fig.update_layout(xaxis_title="Date",yaxis_title="Percentage Cases",title = '%age Confirmed Cases, Negative Cases & Death Cases')
fig.show()
```



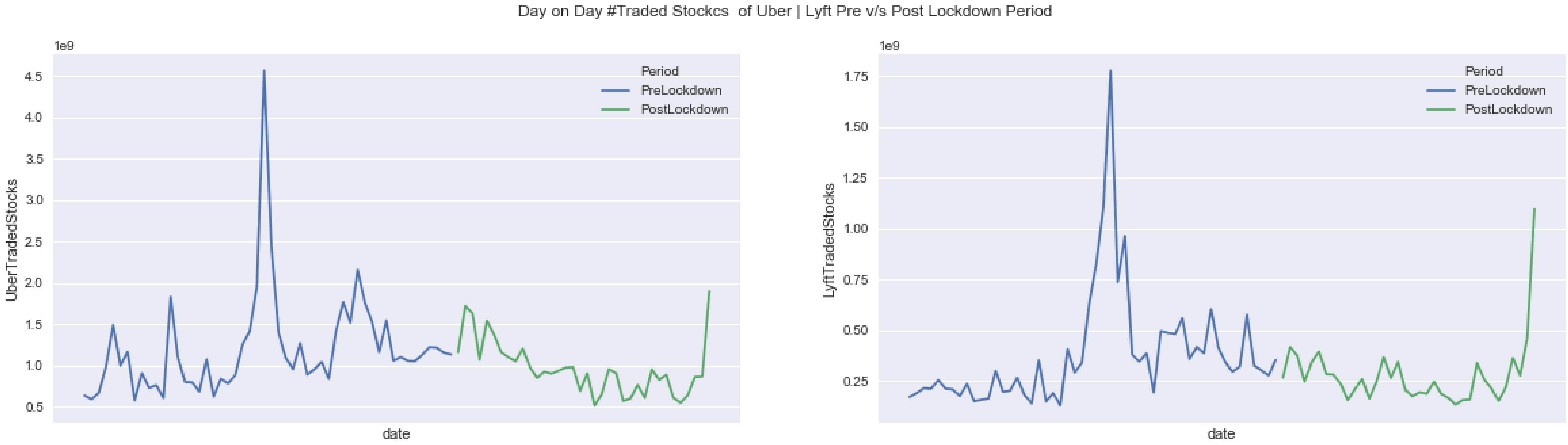
Inference from above graph: This is interesting that in the intial few days of the outbreak there are mostly postive cases, this is due to testing being limited to high potential people whle we can see that with time testing has picked up and negative and positive cases seems to be breaking even in the current scenario and negative cases are more after the complete lockdown, while death rate seems to be gently increasing

Let's Observe Pre v/s Post COVID Outbreak Traded Stocks for Uber/Lyft

In [27]:
```
## Sketch Pre period also for this
lockdown_dt= pd.to_datetime('2020-03-18').strftime('%Y-%m-%d')
x_sel['Period']= np.where(x_sel['date'] >= lockdown_dt, 'PostLockdown', 'PreLockdown')
x_sel['UberTradedStocks']= x_sel['UberVolume']* x_sel['UberClosingPrice']
x_sel['LyftTradedStocks']= x_sel['LyftVolume'] * x_sel['LyftClosingPrice']
```

In [64]:
```
time_eda= pd.to_datetime('2020-01-01').strftime('%Y-%m-%d')
x_tmp= x_sel.copy()
x_tmp = x_tmp[x_tmp['date']>=time_eda]
fig = plt.figure(figsize= (20,5))
plt.subplot(1,2,1)
g =sns.lineplot(x="date", y="UberTradedStocks",hue="Period",data=x_tmp)
g.set(xticks=[])
plt.subplot(1,2,2)
g =sns.lineplot(x="date", y="LyftTradedStocks",hue="Period",data=x_tmp)
g.set(xticks=[])
fig.suptitle("Day on Day #Traded Stockcs  of Uber | Lyft Pre v/s Post Lockdown Period", fontsize=12)
```

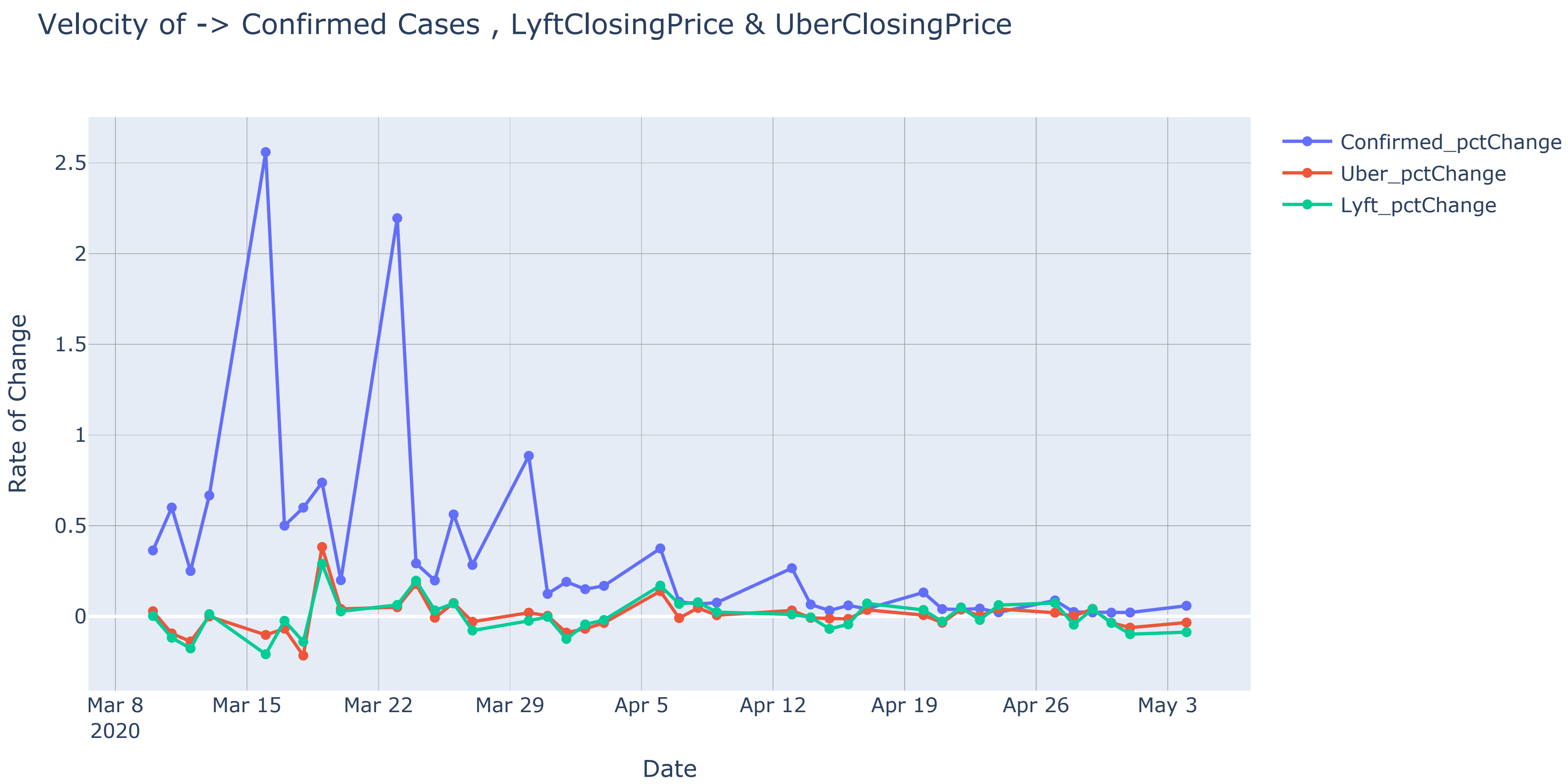
Out[64]:
```
Text(0.5, 0.98, 'Day on Day #Traded Stockcs  of Uber | Lyft Pre v/s Post Lockdown Period')
```



Inference from above graph: We can clearly see that COVID19 outbreak has very badly hit ride sharing market, traded stocks have gone down by very high rate, can be seen from the pre v/s post lockdown period

Let's Plot Precentage Change Day on Day in StockPrices V/s Changes in #Cases

In [29]:
```
df_temp= comb_df.copy()
df_temp['Uber_pctChange'] = df_temp['UberClosingPrice'].pct_change(periods=1)
df_temp['Lyft_pctChange'] = df_temp['LyftClosingPrice'].pct_change(periods=1)
df_temp['Confirmed_pctChange'] = df_temp['cumpositive'].pct_change(periods=1)
df_temp = df_temp.iloc[1:]
fig = go.Figure()
fig.add_trace(go.Scatter(x=df_temp['date'], y=df_temp['Confirmed_pctChange'], mode='lines+markers', name='Confirmed_pctChange'))

fig.add_trace(go.Scatter(x=df_temp['date'], y=df_temp['Uber_pctChange'], mode='lines+markers', name='Uber_pctChange'))

fig.add_trace(go.Scatter(x=df_temp['date'], y=df_temp['Lyft_pctChange'], mode='lines+markers', name='Lyft_pctChange'))
fig.update_layout(xaxis_title="Date",yaxis_title="Rate of Change",
        title = 'Velocity of -> Confirmed Cases , LyftClosingPrice & UberClosingPrice')
fig.show()
```

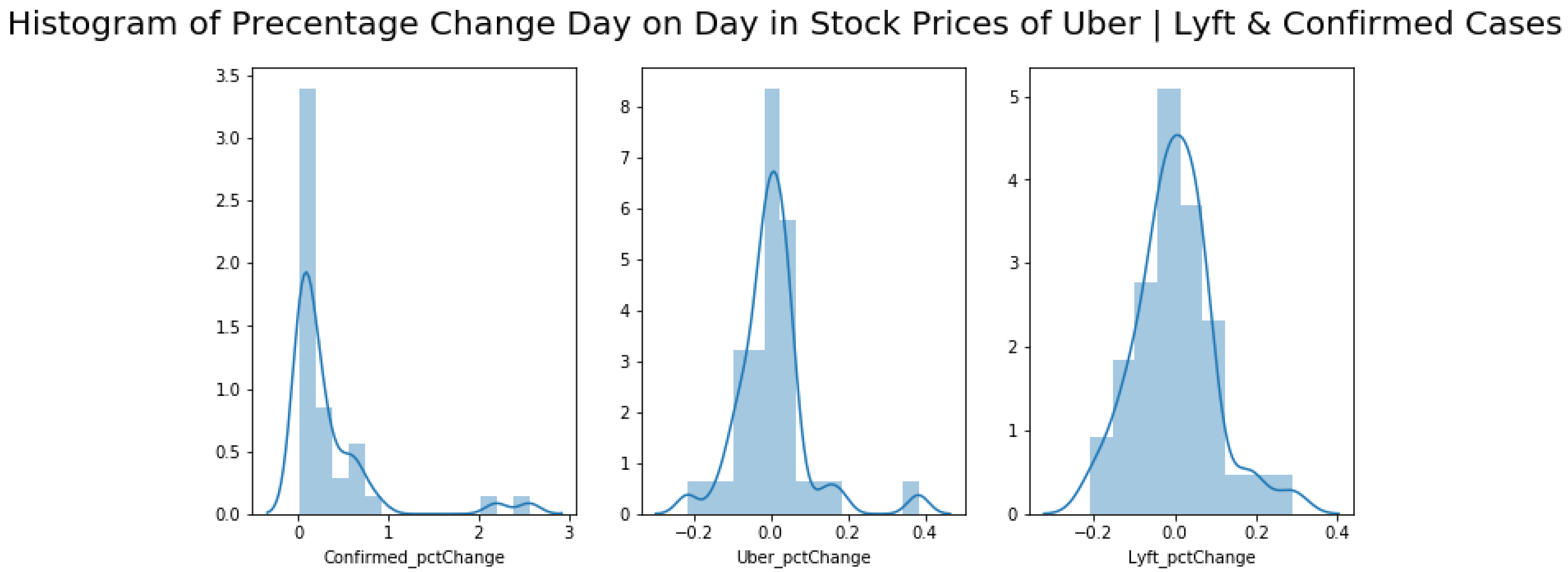


We can draw an inference from above plot is that rate of postive change in confirm case was very high in intital few weeks , later its has come to changes ~20% daily while Uber | Lyft are showing ripple around zero; meaning there are positve and negatve changes as the COVID rates are changing

Let's Plot the Histogram of Percentage Changes to see at what frequency we are observing postive and negative changes

In [30]:
```
#histogram
fig = plt.figure(figsize= (12,5))
plt.subplot(1,3,1)
sns.distplot((df_temp['Confirmed_pctChange']))
plt.subplot(1,3,2)
sns.distplot(df_temp['Uber_pctChange'], label="Uber Changes")
plt.subplot(1,3,3)
sns.distplot((df_temp['Lyft_pctChange']))
fig.suptitle("Histogram of Precentage Change Day on Day in Stock Prices of Uber | Lyft & Confirmed Cases", fontsize=20)
```

Out[30]:
```
Text(0.5, 0.98, 'Histogram of Precentage Change Day on Day in Stock Prices of Uber | Lyft & Confirmed Cases')
```



- ***Inference from above graph: As the velocity in the Positive Cases increases we see that velocity in the Uber & Lyft Price decreases and when the velocity of confirm cases decreases then velocity in the Stock Prices of Uber Lyft Increases***
- ***Changes in the confirmed cases is right skewed which suggests increasing cases while for Uber & Lyft we see that its left skewed which shows a constant decline in this Stock Prices while Lyft has smooth fluctuation***

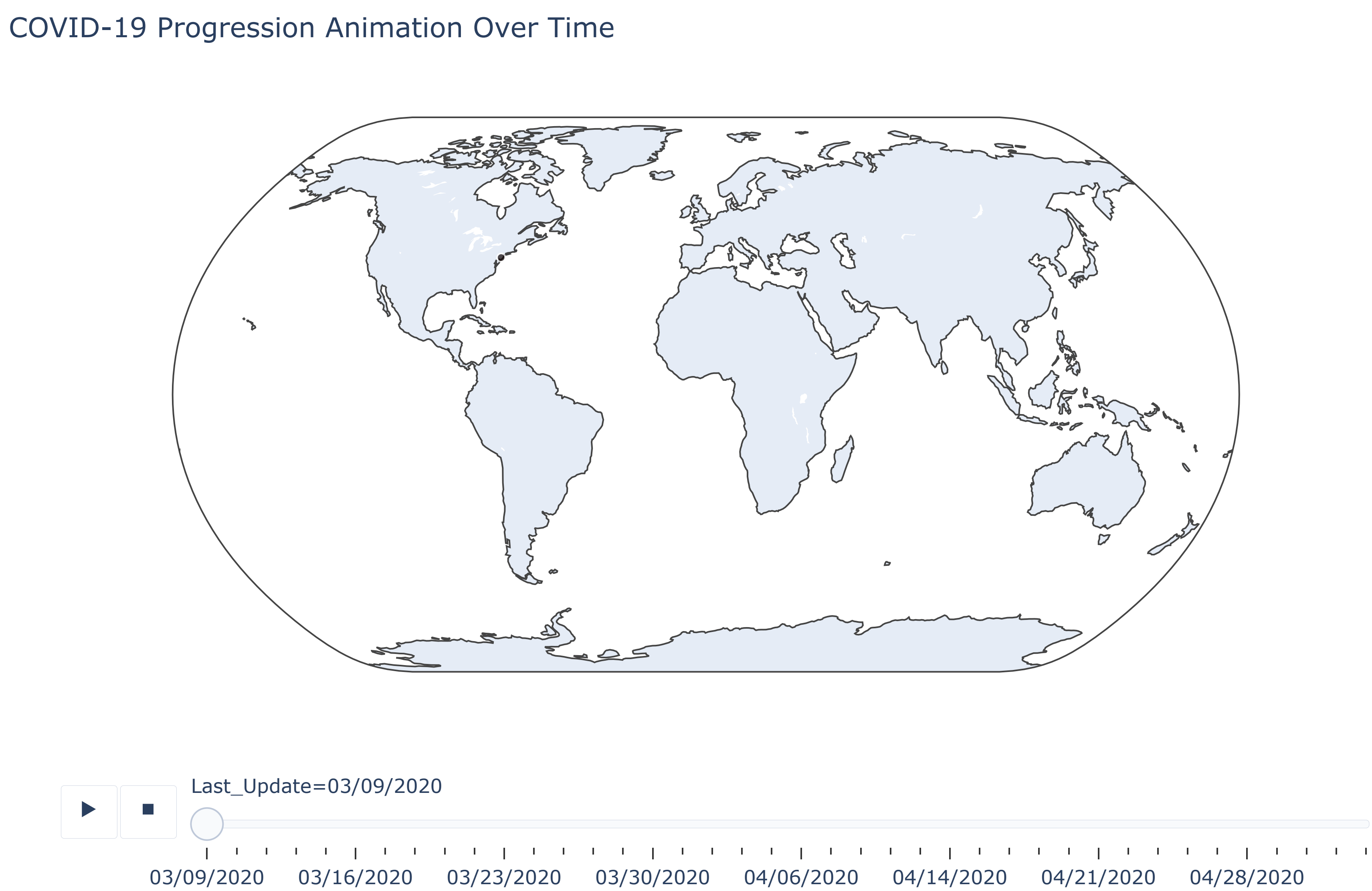
Lets Provide GeoSpatial Mapping of New Jersey COVID Cases with Time

In [31]:
```
df_temp=comb_df.copy()
df_temp['Country_Region']= 'NJ'
df_temp['Lat']= 39.833851
df_temp['Long']= -74.871826

df_temp = df_temp.groupby(['date', 'Country_Region','Lat','Long'])['cumpositive', 'cumdeath'].max().reset_index()
df_temp["date"] = pd.to_datetime(df_temp["date"]).dt.strftime('%m/%d/%Y')
df_temp.columns=['Last_Update','Country_Region','Lat','Long','Confirmed','Deaths']
df_temp['Confirmed'].fillna(0, inplace=True)
df_temp.sort_values('Confirmed', ascending=False).head(3)
```

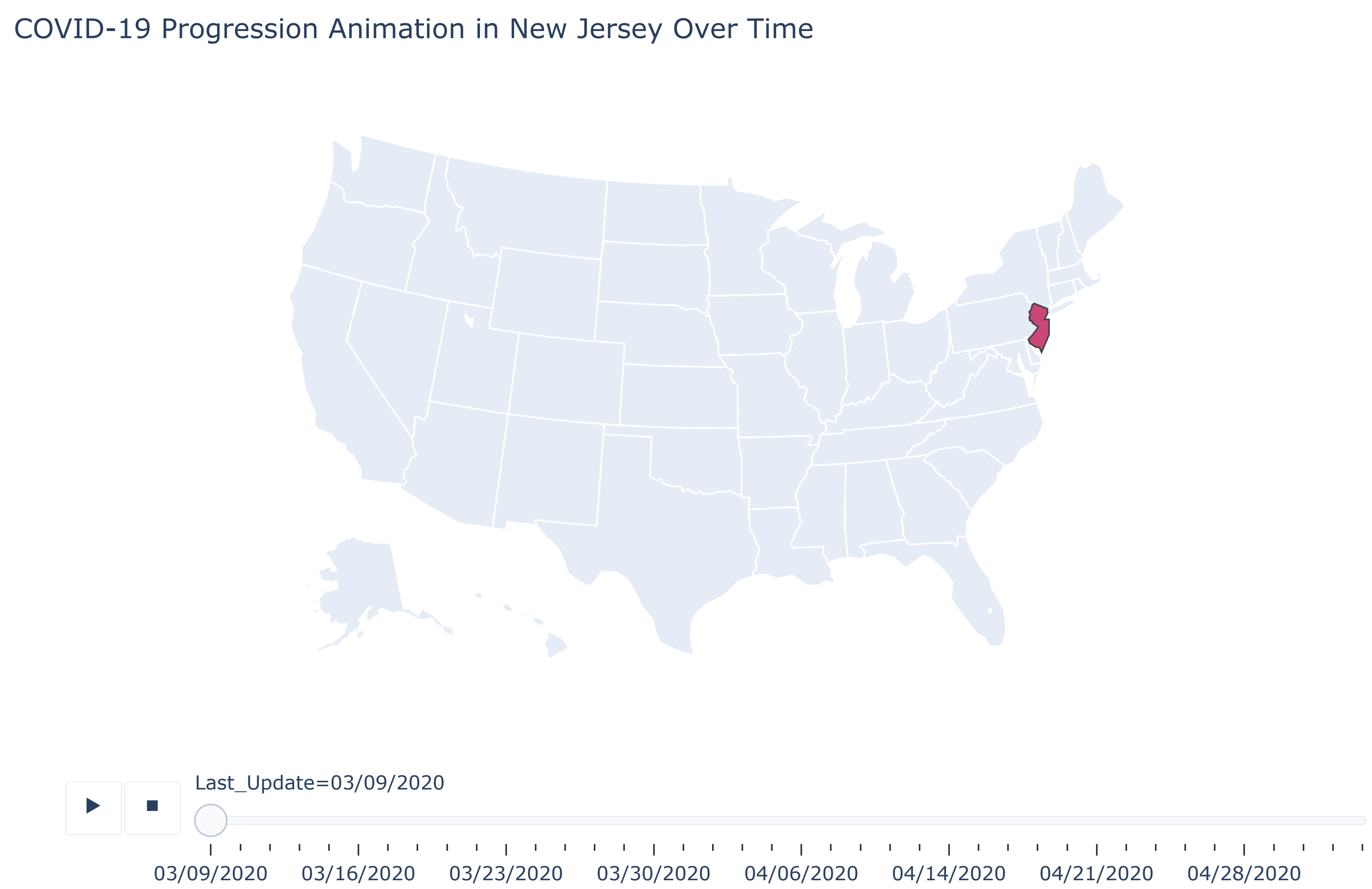
Out[31]:

	Last_Update	Country_Region	Lat	Long	Confirmed	Deaths
39	05/04/2020	NJ	39.833851	-74.871826	128269	7910
38	05/01/2020	NJ	39.833851	-74.871826	121190	7538
37	04/30/2020	NJ	39.833851	-74.871826	118652	7228

In [32]:
```
fig = px.scatter_geo(df_temp,
                     #locations="Country_Region",
                     locationmode='country names',
                     lat='Lat', lon='Long',
                     #hover_name="Country_Region",
                     hover_data=["Confirmed", "Deaths"], animation_frame="Last_Update",
                     color=np.log10(df_temp["Confirmed"]+1)-1, size=np.power(df_temp["Confirmed"]+1, 0.3)-1,
                     range_color= [0, max(np.log10(df_temp["Confirmed"]+1))],
                     title="COVID-19 Progression Animation Over Time",
                     color_continuous_scale=px.colors.sequential.Plasma,
                     projection="natural earth"
                    )
fig.update_coloraxes(colorscale="hot")
fig.update(layout_coloraxis_showscale=False)
fig.show()
```



In [33]:
```
fig = px.choropleth(df_temp,
                    locations="Country_Region",
                    locationmode="USA-states",
                    hover_name="Country_Region",
                    hover_data=["Confirmed", "Deaths"], animation_frame="Last_Update",
                    color=np.log10(df_temp["Confirmed"]),
                    title="COVID-19 Progression Animation in New Jersey Over Time",
                    color_continuous_scale=px.colors.sequential.Plasma,
                    scope="usa",
                   )
fig.update(layout_coloraxis_showscale=False)
fig.show()
```



Part 3: Required Inferences (50%)

3.1 Predicting the COVID19 fatality & #cases over next one week

Use your COVID19 dataset to predict the COVID19 fatality and #cases for the next one week. Use the following four prediction techniques: (i) AR(3), (ii) AR(5), (iii) EWMA with alpha = 0.5, and (iv) EWMA with alpha = 0.8. Make sure that your dataset allows you to verify the one week prediction. For example, use the first three weeks of data to predict the fourth week, and report the accuracy of your predictions using the actual fourth week data. Use metrics learned in class (MAPE as a % and MSE) to report accuracy numbers.

In [34]:

```
ts_data=covid_sel[['date','dailydeath']]
ts_data['WeekNum'] = ((pd.to_datetime(ts_data['date']) - pd.to_datetime(st_dt)).dt.days)//7 +1

posterior_data = ts_data[(ts_data['WeekNum']<=7) & (ts_data['WeekNum']>=4)]
posterior_data = posterior_data.sort_values(by="date").reset_index(drop=True)

weekly_data = ts_data[(ts_data['WeekNum']<=6) & (ts_data['WeekNum']>=4)]
weekly_data = weekly_data.sort_values(by="date").reset_index(drop=True)

test_data = ts_data[(ts_data['WeekNum']==7)]
test_data = test_data.sort_values(by="date").reset_index(drop=True)

print('\033[1m' +'Min Date observed for COVID : ' + str(weekly_data['date'].min()))
print('\033[1m' + 'Max Date observed for COVID: ' + str(weekly_data['date'].max()))

weekly_data['date']=pd.to_datetime(weekly_data['date'])
test_data['date']=pd.to_datetime(test_data['date'])
```

```
Min Date observed for COVID : 2020-03-30
Max Date observed for COVID: 2020-04-19
```

3.1.1 AR(3)

Performing regression Using OLS Method:

In [35]:
```
#Y_hat= B0 + B1(Y_t-1) + B2(Y_t-2) + B3(Y_t-3)
#Predicting #fatalities using AR(3)
# Linear Regression using 3 weeks data to predict 4th weeks' fatalities. Here , n=21 (7 for test data),p=2
def load_data(y_data):
    Y = y_data.to_numpy()   #(21,)
    Y=Y.reshape(-1,1)     #(21,1)
    return Y

def get_beta_coeff(Y,p):
    low=0
    high=p
    Y_row=Y.T
    Y_row.tolist()
    Y_row = Y_row[0]

    ones=[1]
    d = []
    while high < len(Y_row):
        temp=[*ones,*Y_row[low: high]]
        d.append(temp)
        low += 1
        high += 1

    X=np.asarray(d)    #(18,4)
    X_Transpose=X.T               #(4,18)
    XT_X=np.dot(X_Transpose,X)    #(4,4)
    inv= np.linalg.inv(XT_X)  #(4,4)

    beta_OLS = np.dot(np.dot(inv,X_Transpose), Y[p:len(Y)])   #(18,1)
    return beta_OLS,Y

def predict(beta_coeff,Y,p):
    for i in range(7):
        f = Y[len(Y)-p:]
        f = f.T
        f = f[0].tolist()
        f.insert(0, 1)
        f=np.asarray(f)
        f=f.reshape(-1,p+1)
        Y=np.concatenate((Y,np.dot(f,beta_coeff)))
        beta_coeff,Y=get_beta_coeff(Y,p)
    return Y

def compare_y(true_data,pred_data):
    true_y=true_data['dailydeath'][-7:]
    predicted_y=pred_data[-7:]
    pred_y=[j for sub in predicted_y for j in sub]

    #Comparison b/w True and Predicted values
    table = pd.DataFrame(columns=['True Value','Predicted Value'])
    table['True Value']=true_y
    table['Predicted Value']=pred_y
    print(table)
    return true_y,pred_y

def get_accuracy(true_y,pred_y):
    # MSE = (Y[-7:]-test_data['dailydeath'])/100
        mse=np.mean((true_y-pred_y)**2)
        print('\033[1m' + "Mean Squared Error is :",mse)
```

```
    #MAPE calculation as a % | Formula: 1/n Summation(|(true-predicted)/true|*100)
        pred_y = np.round(pred_y)
        mape=np.sum(np.abs((true_y-pred_y)/true_y))/7
        print('\033[1m' + "MAPE as a %:",mape*100)
```

In [36]:
```
def AR(p):
    y_data = load_data(weekly_data['dailydeath'])
    beta_OLS,Y = get_beta_coeff(y_data,p)
    pred_data = predict(beta_OLS,Y,p)
    true_y,pred_y = compare_y(test_data,pred_data)
    get_accuracy(true_y,pred_y)
    return true_y,pred_y
```

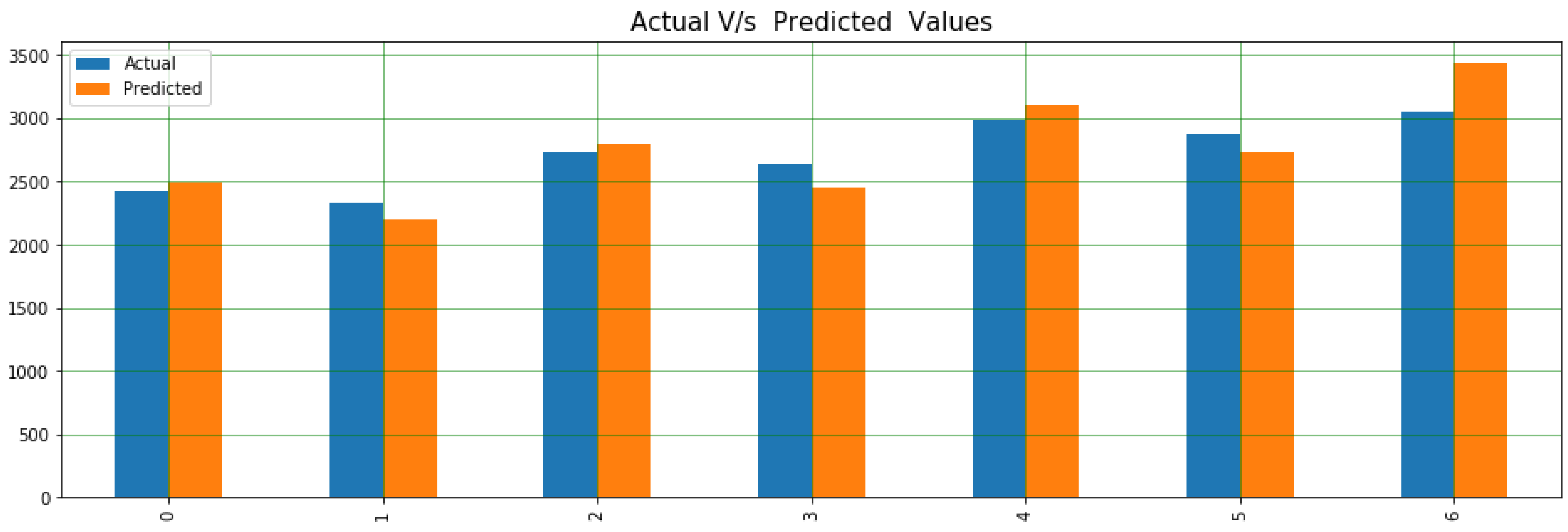
In [37]:
```
def plot_bar_actual_pred(test_data,predicted_data, title):
  var= title
  plt.plot(test_data, predicted_data)
  plt.title(var, size=15)
  plt.xlabel('Actual', size= 15)
  plt.ylabel('Predicted', size=15)
  plt.show()
  print()
```

In [38]:
```
def plot_actual_predicted(test_data, predicted_data):
  y_test_flat= test_data
  y_pred_flat=predicted_data

  df = pd.DataFrame({'Actual': y_test_flat, 'Predicted': y_pred_flat})
  df1 = df.head(25)
  df1.plot(kind='bar',figsize=(16,5))
  plt.grid(which='major', linestyle='-', linewidth='0.5', color='green')
  plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')
  plt.title('Actual V/s  Predicted  Values',size=15)
  plt.show()
```

Output for AR(p=3)

In [39]:
```
true_y,pred_y= AR(p=3)
print('\n')
# plot_bar_actual_pred(true_y,pred_y,'Actual v/s Predicted for AR(p=3)')
plot_actual_predicted(true_y, pred_y)
```

```
   True Value  Predicted Value
0        2422      2496.838311
1        2331      2195.225901
2        2732      2793.633614
3        2636      2457.117042
4        2981      3109.581806
5        2882      2731.678019
6        3056      3442.454922
```
Mean Squared Error is : 35472.93988355075
MAPE as a %: 5.736289660514681



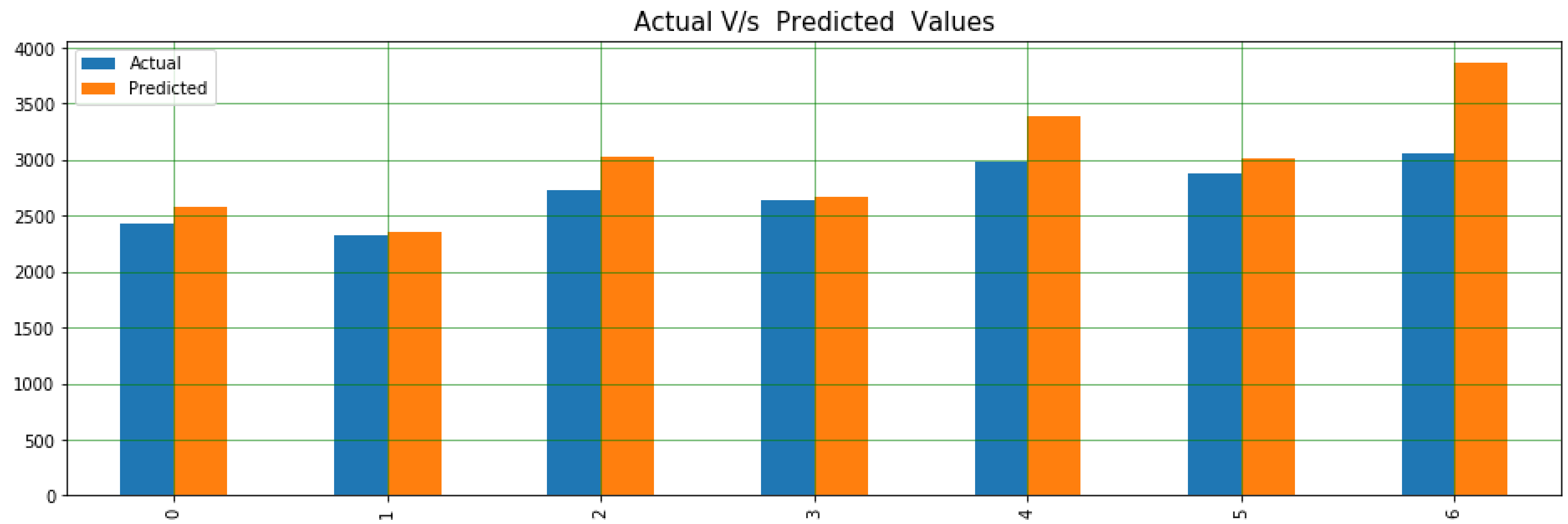
3.1.2 AR(5)

Output for AR(p=5)

In [40]:
```
true_y,pred_y= AR(p=5)
print('\n')
# plot_bar_actual_pred(true_y,pred_y,'Actual v/s Predicted for AR(p=5)')
plot_actual_predicted(true_y, pred_y)
```

```
   True Value  Predicted Value
0        2422      2578.721039
1        2331      2348.936320
2        2732      3023.222311
3        2636      2668.866613
4        2981      3392.347771
5        2882      3017.146203
6        3056      3871.760795
```

Mean Squared Error is : 137673.00060242004
MAPE as a %: 9.190145488732558



3.1.3 EWMA with alpha = 0.5

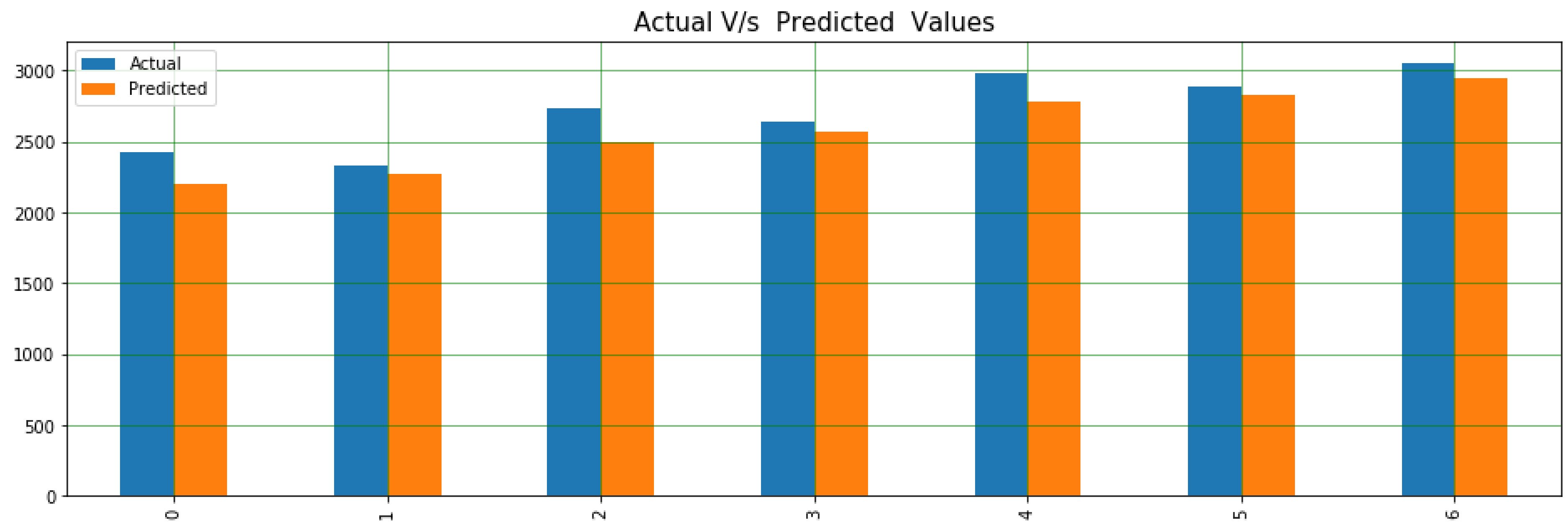
In [41]:
```
def exponential_smoothing(train, alpha, test):
    """given a series and alpha, return series of expoentially smoothed points"""
    results = np.zeros_like(train)

    # first value remains the same as series,
    # as there is no history to learn from
    results[0] = train[0]
    for t in range(1, train.shape[0]):
        results[t] = alpha * train[t] + (1 - alpha) * results[t - 1]

    ans = np.zeros_like(test)
    ans[0]= results[20] * (1 - alpha) + alpha * test[0]
    for t in range(1, test.shape[0]):
        ans[t] = alpha * test[t] + (1 - alpha) * ans[t - 1]

    return ans
```

In [42]:
```
def compare(EMA_predicted,test_data):
    table=pd.DataFrame(columns=['true_values','prediction'])
    # print("table",table)
    table['prediction'] = EMA_predicted
    table['true_values'] = test_data['dailydeath']
    print(table)
    true_y = test_data['dailydeath']
    pred_y = EMA_predicted
    mse=np.mean((true_y-pred_y)**2)
    print('\033[1m' + "Mean Squared Error is :",mse)

    #MAPE calculation as a % | Formula: 1/n Summation(|(true-predicted)/true|*100)
    pred_y = np.round(pred_y)
    mape=np.sum(np.abs((true_y-pred_y)/true_y))/7
    print('\033[1m' + "MAPE as a %:",mape*100)
```

In [43]:
```
EMA_predicted= exponential_smoothing(weekly_data['dailydeath'], 0.5, test_data['dailydeath'])
estimated_values=test_data['dailydeath'].copy() # replace testdata with your test dataset
estimated_values['predict'] = EMA_predicted[1:]
```

In [44]:
```
compare(EMA_predicted,test_data)
print('\n')
# plot_bar_actual_pred(test_data['dailydeath'],EMA_predicted,'Actual v/s Predicted for EWMA (alpha = 0.5)')
plot_actual_predicted(list(test_data['dailydeath']),list(EMA_predicted))
```

```
   true_values  prediction
0         2422        2202
1         2331        2266
2         2732        2499
3         2636        2567
4         2981        2774
5         2882        2828
6         3056        2942
```
Mean Squared Error is : 24348.0
MAPE as a %: 5.080871674137092

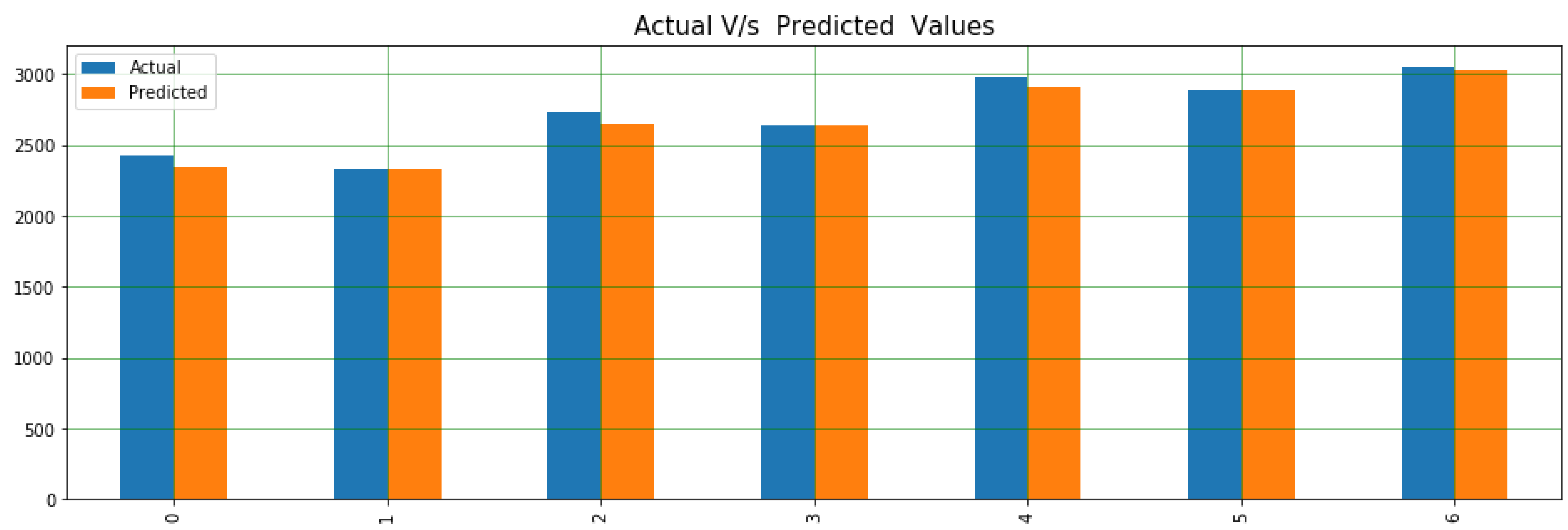


3.1.4 EWMA with alpha = 0.8

In [45]:
```
EMA_predicted= exponential_smoothing(weekly_data['dailydeath'], 0.8, test_data['dailydeath'])
estimated_values=test_data['dailydeath'].copy() # replace testdata with your test dataset
estimated_values['predict'] = EMA_predicted[1:]
```

In [46]:
```
compare(EMA_predicted,test_data)
print('\n')
# plot_bar_actual_pred(test_data['dailydeath'],EMA_predicted,'Actual v/s Predicted for EWMA (alpha = 0.8)')
plot_actual_predicted(list(test_data['dailydeath']),list(EMA_predicted))
```

```
   true_values  prediction
0         2422        2337
1         2331        2332
2         2732        2652
3         2636        2639
4         2981        2912
5         2882        2888
6         3056        3022
```
Mean Squared Error is : 2798.285714285714
MAPE as a %: 1.4614109325597018



Inferences:

- With AR(p=3) and AR(p=5) ....
- With EWMA(alpha =0.5) and EWMA(alpha =0.8) ...

3.2 Apply the Wald's test, Z-test, and t-test to check whether the mean of COVID19 deaths and #cases are different from the first week to the last week

Apply the Wald's test, Z-test, and t-test (assume all are applicable) to check whether the mean of COVID19 deaths and #cases are different from the first week to the last week in your dataset. Use MLE for Wald's test as the estimator. Note, you have to report results for deaths and #cases separately, so think of this as two inferences. After running the test and reporting the numbers, check and comment on whether the tests are applicable or not. First use one-sample tests by computing the mean of the first week data and using that as guess for last week data. Then, repeat with a two-sample version of Wald and t-tests. For t-test, use both paired and unpaired tests. Use alpha value of 0.05 for all. For t-test, the threshold to check against is tn-1, alpha/2 for two-tailed and tn-1, alpha for one-tailed, where n is the number of data points. You can find these values in online t tables, similar to z tables.

3.2.1 Use MLE for Wald's test as the estimator

In [ ]:

In [ ]:

In [ ]:

3.2.2 Two-sample version of Wald and t-tests

In [ ]:

In [ ]:

In [ ]:

3.2.3 Z-test

In [ ]:

In [ ]:

In [ ]:

3.3 Equality of distributions (distribution of first week and last week), using K-S test and Permutation test

Repeat inference 2 above but for equality of distributions (distribution of first week and last week), using K-S test and Permutation test. For the K-S test, use both 1-sample and 2-sample tests. For the 1-sample test, try Poisson, Geometric, and Binomial. To obtain parameters of these distributions to check against in 1-sample KS, use MME on first week's data to obtain parameters of the distribution, and then check whether the last week's data has the distribution with the obtained MME parameters. Use a threshold of 0.05 for both K-S test and Permutation test.

3.3.1 K-S Test

In [ ]:

In [ ]:

In [ ]:

3.3.2 Permutation Test

In [ ]:

In [ ]:

In [ ]:

3.4 Pearson correlation for #deaths and Total Traded Stocks, #cases and Total Traded Stocks

Report the Pearson correlation value for #deaths and your X dataset, and also for #cases and your X dataset over one month of data. Use the most relevant column in X to compare against the covid numbers.

In [47]:
```
import math
def p_coeff(a,b):

    ab_n1 = 0
    ab_d1 = 0
    ab_d2 = 0

    mean_a = sum(a)/len(a)
    mean_b = sum(b)/len(b)
    for i, j in zip(a,b):
        ab_n1 += (i- mean_a) * (j- mean_b)
        ab_d1 += (i- mean_a) * (i- mean_a)
        ab_d2 += (j- mean_b) * (j- mean_b)
    ab = ab_n1 / (math.sqrt(ab_d1) * math.sqrt(ab_d2))
    return ab
```

Calculating Total Traded Stocks for the Day

In [48]:
```
comb_df['UberTradedStocks']= comb_df['UberVolume']* comb_df['UberClosingPrice']
comb_df['LyftTradedStocks']= comb_df['LyftVolume'] * comb_df['LyftClosingPrice']
```

3.4.1 Pearson correlation for #deaths and Total Traded Stocks

In [49]:
```
corr= p_coeff(comb_df['cumdeath'], comb_df['UberTradedStocks'])
print('\033[1m' + 'Pearsons correlation of #deaths and Stock Price of Uber: %.3f' % corr)

corr = p_coeff(comb_df['cumdeath'], comb_df['LyftTradedStocks'])
print('\033[1m' + 'Pearsons correlation of #deaths and Stock Price of Lyft: %.3f' % corr)
```

```
Pearsons correlation of #deaths and Stock Price of Uber: -0.719
Pearsons correlation of #deaths and Stock Price of Lyft: -0.417
```

Inference: We can observe a high -ve linear correlation (-0.72) between stock prices of Uber/Lyft v/s the Deaths, this means that increase in #deaths day on day has adversely affected ride sharing company with less people moving out

3.4.2 Pearson correlation for #cases and Stock Price

In [50]:
```
corr= p_coeff(comb_df['cumpositive'],  comb_df['UberTradedStocks'])
print('\033[1m' + 'Pearsons correlation of #Confirmed Cases and Stock Price of Uber: %.3f' % corr)

corr= p_coeff(comb_df['cumpositive'], comb_df['LyftTradedStocks'])
print('\033[1m' + 'Pearsons correlation of #Confirmed Cases and Stock Price of Lyft: %.3f' % corr)
```

```
Pearsons correlation of #Confirmed Cases and Stock Price of Uber: -0.773
Pearsons correlation of #Confirmed Cases and Stock Price of Lyft: -0.487
```

Inference: We can observe a high -ve (-0.78) linear correlation between stock prices of Uber/Lyft v/s the #Confirm cases, this means that increase in #Confim cases day on day has brought the city to a halt and ride sharing company stocks are going down as less and less people are moving out

3.5 Posterior Distributions for daily deaths parameter estimator

Assume the daily deaths are Poisson distributed with parameter lambda. Assume an Exponential prior (with mean beta) on lambda. To find beta for the prior, equate the mean of the Exponential prior to that of the Poisson lambda_MME. That is, find the MME of lambda using the first week's data, and equate this lambda to the mean of Exp(1/beta) to find beta for the prior. Use first week's data to obtain the posterior for lambda via Bayesian inference. Now, use second week's data to obtain the new posterior, using prior as posterior after week 1. Repeat till the end of week 4. Plot all posterior distributions on one graph. Report the MAP for all posteriors.

Posterior becomes a Gamma Distribution with params (Summ(x_i)+1,n + 1/beta)

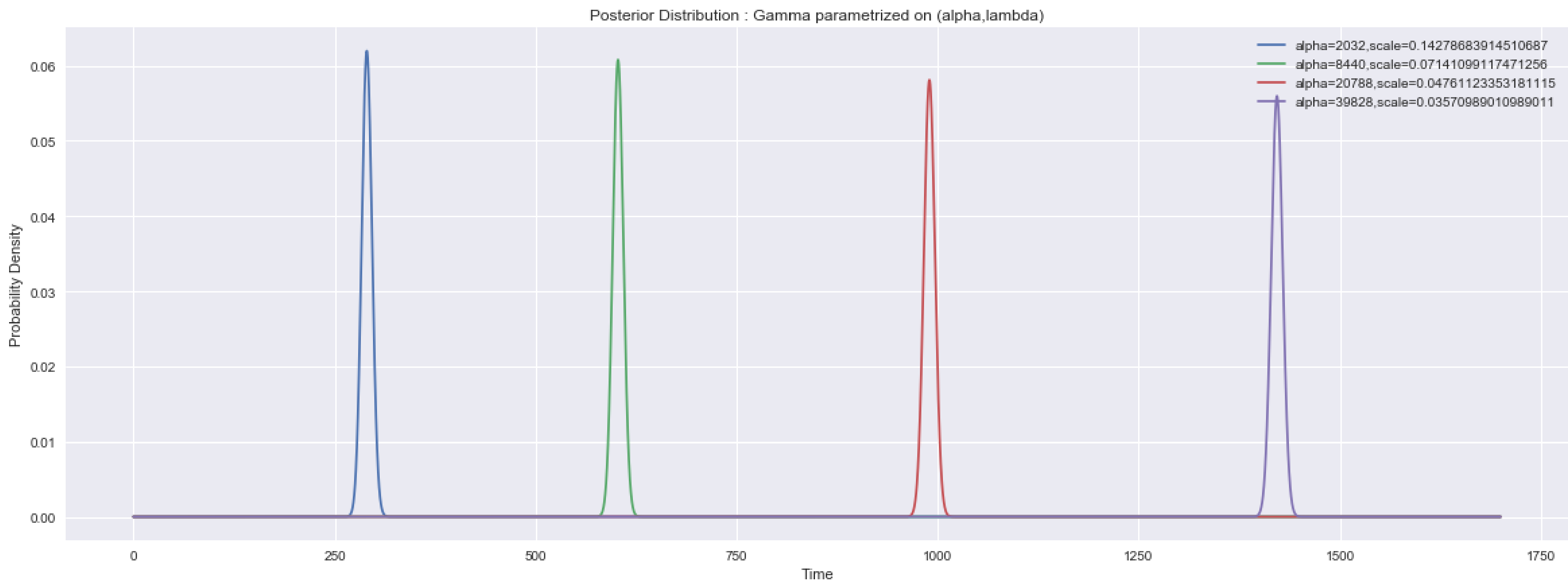
In [51]:
```
import numpy as np
from scipy.stats import gamma
import matplotlib.pyplot as plt

plt.style.use('seaborn')
fig_size = plt.rcParams["figure.figsize"]
fig_size[0] = 20
fig_size[1] = 7
plt.rcParams["figure.figsize"] = fig_size

global first_x

def get_first_x():
    weekwise = np.array_split(posterior_data['dailydeath'], 4)
    first_x=np.sum(weekwise[0])
    return first_x


def get_posterior(week_num, sum_x):
    first_x=get_first_x()
    x = np.linspace(0,1700, 1000)
    n = week_num*7
    alpha= sum_x +1
    lambda_= n+(7/first_x)

    print('\033[1m' + "MAP for Week: {0} = {1}".format(week_num,alpha/lambda_))
    return alpha,lambda_


def plot_posterior(alpha,lambda_):
    x = np.linspace(0,1700, 10000)
    scale= 1/lambda_
    res = gamma.pdf(x, alpha, scale=1/lambda_)
    label = "alpha={0},scale={1}".format(alpha, scale)

    title = "Posterior Distribution : Gamma parametrized on (alpha,lambda)"
    plt.title(title)
    plt.xlabel("Time")
    plt.ylabel("Probability Density")
    plt.plot(x, res,label=label)
```

Report MAP and Plot all posterior distributions on one graph

In [52]:
```
def init_data():
    weekwise = np.array_split(posterior_data['dailydeath'], 4)
    rolling_sum=0
    cumsum_weekwise=[]
    for i in range(4):
        rolling_sum=rolling_sum+np.sum(weekwise[i])
        cumsum_weekwise.append(rolling_sum)
        alpha,lambda_= get_posterior(i+1,cumsum_weekwise[i])
        print('\033[1m' + "Posterior Params for Week: {0} are alpha = {1} and lambda = {2}\n".format(i+1,alpha,lambda_))
        plot_posterior(alpha,lambda_)
        plt.legend(loc="upper right")
```

In [53]:
```
init_data()
```

```
MAP for Week: 1 = 290.14285714285717
Posterior Params for Week: 1 are alpha = 2032 and lambda = 7.003446578040374

MAP for Week: 2 = 602.708765514574
Posterior Params for Week: 2 are alpha = 8440 and lambda = 14.003446578040375

MAP for Week: 3 = 989.7423226592902
Posterior Params for Week: 3 are alpha = 20788 and lambda = 21.003446578040375

MAP for Week: 4 = 1422.2535032967032
Posterior Params for Week: 4 are alpha = 39828 and lambda = 28.003446578040375
```



Part 4: Creative Inferences (30%)

Propose three new inferences for your dataset and solve them using tools learned in class. You will be graded on creativity/practicality of your inferences. For each inference you propose, provide a paragraph of text to explain why this inference is practical and useful. Also comment on the results of your inference, as appropriate. See “Sample inferences section below for ideas. Only use tools/tests learned in class. This will be 30% of the project grade.

Hypothesis1: Performing Chi-Square test to show due to Uber being functional Covid Spread Quickly and once they were shut spread went down

Using Chi-square independence test to check if Uber Stock Prices impacted COVID19 cases

Step 1: Define the Hypothesis

For this we will be creating two lables for COVID19 changes in Confirmed Cases ("Positive_pctChange") as postive and negative , and similarly changes in Closing price for Uber ("Uber_pctChange") as positive and negative

For our example, the hypothesis are:

- ***H0: The Change in Confirmed Cases(Positive_pctChange) and changes in Closing price for Uber ("Uber_pctChange") are independent (which means they are not associated)***
- **H1: Change in Confirmed Cases and changes in Closing price for Uber are not independent (which means they are associated)**

In [54]:
```
var1= 'UberClosingPrice'
var2= 'cumpositive'

comb_df['Uber_pctChange'] = comb_df[var1].pct_change(periods=1)
comb_df['Confirmed_pctChange'] = comb_df[var2].pct_change(periods=1)
comb_df=comb_df.iloc[1:]

comb_df['Uber_Slope'] = comb_df['Uber_pctChange'].pct_change(periods=1)
comb_df['Confirmed_Slope'] = comb_df['Confirmed_pctChange'].pct_change(periods=1)
comb_df=comb_df.iloc[1:]
```

Crating Lables for Changes in Confirm Cases and Uber's Closing Price Day On Day

In [55]:
```
comb_df['Confirmed_Label']= np.where(comb_df['Confirmed_Slope'] >= 0, 'Positive', 'Negative')
comb_df['Uber_Label']= np.where(comb_df['Uber_Slope'] >= 0, 'Positive', 'Negative')
```

In [56]:
```
comb_df.iloc[:,20:28].head(2)
```

Out[56]:

	Uber_pctChange	Confirmed_pctChange	Uber_Slope	Confirmed_Slope	Confirmed_Label	Uber_Label
40	-0.094235	0.60	-4.318268	0.650000	Positive	Negative
39	-0.138338	0.25	0.468009	-0.583333	Negative	Positive

Step2: Choose a significance Level

For the null hypothesis to be rejected the p-value should be less than the significance level.

Lower α values are generally preferred which may be in the range of 0.01 to 0.10.We choose α = 0.05

Step3: Create Contingency table

In [57]:
```
Q=pd.crosstab(comb_df['Confirmed_Label'], comb_df['Uber_Label'], rownames=['Confirmed_Label'], colnames=['Uber_Label'])
print(Q)

Q_table = comb_df.groupby(['Confirmed_Label','Uber_Label'])['date'].count()
Q_table = Q_table.reset_index()
Q_table.columns = ['Confirmed_Label','Uber_Label','TotalDays']
```

```
Uber_Label       Negative  Positive
Confirmed_Label
Negative               14         5
Positive               13         6
```

Step4: Calculate Expected Frequency

In [58]:
```
comb_df.shape
total=  Q_table['TotalDays'].sum()

per_cp= round(Q_table[(Q_table['Confirmed_Label']== 'Positive')].TotalDays.sum()/total,2)
per_up= round(Q_table[(Q_table['Uber_Label']== 'Positive')].TotalDays.sum()/total,2)

ob_cp_up= Q_table[(Q_table['Confirmed_Label']== 'Positive') & (Q_table['Uber_Label'] =='Positive')].TotalDays.sum()
ob_cp_un= Q_table[(Q_table['Confirmed_Label']== 'Positive') & (Q_table['Uber_Label'] =='Negative')].TotalDays.sum()
ob_cn_up= Q_table[(Q_table['Confirmed_Label']== 'Negative') & (Q_table['Uber_Label'] =='Positive')].TotalDays.sum()
ob_cn_un= Q_table[(Q_table['Confirmed_Label']== 'Negative') & (Q_table['Uber_Label'] =='Negative')].TotalDays.sum()

ex_cp_up= per_cp*per_up*total
ex_cp_un= per_cp*(1-per_up)*total
ex_cn_up= (1-per_cp)*per_up*total
ex_cn_un= (1-per_cp)*(1-per_up)*total


print(total, per_cp, per_up, ob_cp_up, ob_cp_un, ob_cn_up, ob_cn_un, ex_cp_up, ex_cp_un, ex_cn_up, ex_cn_un)
```

```
38 0.5 0.29 6 13 5 14 5.51 13.489999999999998 5.51 13.489999999999998
```

Step5: Calculate Chi-Square Statistic

In [59]:
```
def diff_sq(Obs, Exp):
    return ((Obs-Exp)**2)/Exp
```

In [60]:
```
Q= diff_sq(ob_cp_up, ex_cp_up) + diff_sq(ob_cp_un, ex_cp_un) + diff_sq(ob_cn_up, ex_cn_up) + diff_sq(ob_cn_un, ex_cn_un)

print('\033[1m' + 'Q statistics value: ' + str(Q))
```

Q statistics value: 0.12785971728739043

Step6: Calculate degrees of freedom

In [61]:
```
total_rows=2
total_cols=2
dfr = (total_rows - 1) * (total_cols - 1)
print('\033[1m' + 'degree of freedom: ' + str(dfr))
```

```
degree of freedom: 1
```

Step7: Find p-value

calculate the p-value from this website: https://www.socscistatistics.com/pvalues/chidistribution.aspx (https://www.socscistatistics.com/pvalues/chidistribution.aspx)

In [62]:
```
pval=.720724
```

In [63]:
```
# select significance value
alpha = 0.05
# Determine whether to reject or keep your null hypothesis
print('\033[1m' +  'significance=%.3f, p=%.3f' % (alpha, pval))
if pval <= alpha:
    print('\033[1m' +  'COVID spread due to Uber being functinal are associated (reject H0)')
else:
    print('\033[1m' +  'COVID spread due to Uber being functinal are not associated(fail to reject H0)')
```

```
significance=0.050, p=0.721
COVID spread due to Uber being functinal are not associated(fail to reject H0)
```

Inference1: Below are the inference for H1

- ***We Observe that the Null Hypotheiss that the COVID Spread due to Uber being funcitonal are not associated, hence we fail to reject H0***
- **For our example we took alpha = 0.05 but p-val is not statiscally significant with value 0.721 so we fail to reject our Null hypothesis**

Hypothesis2: Using K-S Test to show that COVID Positive Cases fluctuation and Uber Stock fluctuation follows the Same distribution

In [ ]:

In [ ]:

In [ ]:

In [ ]:

Inference2: Below are the inference for H2

In [ ]:

In [ ]:

In [ ]:

In [ ]:

Inference3: Linear regression to find the impact on Stock Prices of Uber +Lyft because of the severity of covid19 duration, feature as (+ve -ve death), fetching predicted covid values of (+ve -ve death) from Part 3.1

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: