# Final Mark: 9 / 10

1. Relevant Dataset and proper visualization methodologies were used
2. EDA was helpful for the understanding the data
3. Linear Regression and Decision Tree are used and results are compared
4. The conclusion and insight was reasonable based on the result from the analysis

# The US Stock Market Reaction to COVID-19 Vaccine News

## I. Problem Statement / Research Topic

The research topic is on analysing the role which COVID-19 vaccine related news and sentiment impacts the performance of various stocks on the United States Stock exchange. We sought to research this topic as the stock market is a useful barometer to assess business confidence across the entire economy and specific industries.

Predicting the relationship and correlation between vaccine related news and changes in the stock market is useful both in predicting future trends which might occur as vaccines for COVID-19 are developed and released.

Having lived through the rapid changes brought on by the COVID-19 outbreak and with a keen interest in financial markets, we hope that this project can help shed new insight into the role which news and sentiment play in impacting financial markets.

## II. Dataset and Data Preparation

The first dataset is a dataset which measures posts on the social media site Twitter with hashtags relating to the COVID-19 vaccine development. This can be used to judge whenever news or information is revealed regarding developments of COVID-19 vaccines and the public's response to them. It can be accessed at the link below

https://www.kaggle.com/ritesh2000/covid19-vaccine-tweets (https://www.kaggle.com/ritesh2000/covid19-vaccine-tweets)

The second dataset we are using is a data on the price of the Dow Jones Industrial Average; a popular stock market index including the largest corporations in the United States. It can be accessed below.

https://finance.yahoo.com/quote/%5EDJI/history?period1=1577836800&period2=1606694400&interval=1d&filter=history&frequency=1d&includeAdjustedClose=(https://finance.yahoo.com/quote/%5EDJI/history?period1=1577836800&period2=1606694400&interval=1d&filter=history&frequency=1d&includeAdjustedClose=

In [1]:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import datetime
import re
import random
```

## COVID-19 Vaccine Tweets Dataset

The dataset consists of **Tweets with "#Covid19" Vaccine Hashtags** that were collected from India throughout the year of 2020. In this dataset, we are going to perform text mining and analysis via Natural Language Processing (NLP) and find quality insights from unstructured textual data from unstructured textual data.

The following steps were executed for this dataset:

1. Simple Cleaning
2. Tokenization
3. Stemmatisation
4. Stop Word Removal

In [4]:

```
tweets_df = pd.read_csv('file.csv')

#Data Cleaning for Covid-19 Tweets Dataset

#Filter & remove unneccesary data in the dataset: #remove non-eng tweet & remove @usern
ames
columns_needed= ['date','username','tweet']

tweets_df = tweets_df[columns_needed]

# Simple Cleaning
def date_cleaned(date):
    date = pd.to_datetime(date, format='%Y-%m-%d')
    return date

tweets_df['date'] = tweets_df['date'].apply(date_cleaned)

def text_clean(tweet):
    tweet = re.sub('[_,$&!;%]', '', tweet)
    tweet = re.sub('[^0-9a-zA-Z\ ]', '', tweet)
    tweet = tweet.lower()
    tweet = tweet.strip()
    return tweet

tweets_df['tweet_cleaned'] = tweets_df['tweet'].apply(text_clean)

# Tokenisation & Stemmatisation

from nltk.tokenize import word_tokenize
from nltk.stem import PorterStemmer

def stemmatisation(x):

        desc_token = word_tokenize(x)

        stemmer = PorterStemmer()

        stemmed = []

        for word in desc_token:
            stemmed.append(stemmer.stem(word))

        stemmed_str = ' '.join(map(str,stemmed))

        return stemmed_str

tweets_df['tweet_stem'] = tweets_df['tweet_cleaned'].apply(stemmatisation)

tweets_df.head()

# Stop Word Removal

from nltk.corpus import stopwords

STOP_WORDS = stopwords.words('english')

def cleaned(x):

    desc_token = word_tokenize(x)
```

```
    stemmed_no_stopwords = []

    for word in desc_token:
        if word not in STOP_WORDS:
            stemmed_no_stopwords.append(word)

    stemmed_no_stopwords_str = ' '.join(map(str,stemmed_no_stopwords))

    return stemmed_no_stopwords_str

tweets_df['tweet_final'] = tweets_df['tweet_stem'].apply(cleaned)

tweets_df.head()
```

Out[4]:

| | date | username | tweet | tweet_cleaned | tweet_stem | tweet_fi |
|---|---|---|---|---|---|---|
| **0** | 2020-10-22 | to_fly_to_live | @ANI Isn't it the best poll promise ever?? Fre... | ani isnt it the best poll promise ever free co... | ani isnt it the best poll promis ever free cov... | ani isnt best promis ever f covid vac |
| **1** | 2020-10-22 | utkarshsinha07 | Now states shall have wait for thier Vidhan Sa... | now states shall have wait for thier vidhan sa... | now state shall have wait for thier vidhan sab... | state shall wait th vidhan sabha el ge |
| **2** | 2020-10-22 | batolebazi | जिस मदारी ने ट्रेन तक नहीं चलाई और तुम पत्नी व... | biharpolls covidvaccine httpstco5cshy52bvz | biharpol covidvaccin httpstco5cshy52bvz | bihar covidvac httpstco5cshy52 |
| **3** | 2020-10-22 | bak_sahil | @MisseeMonis They said vaccine for all but not... | misseemonis they said vaccine for all but not ... | misseemoni they said vaccin for all but not wh... | misseemoni s vaccin free cc vaccin new |
| **4** | 2020-10-22 | ivibhatweedy | BJP really presenting "free COVID vaccine" as ... | bjp really presenting free covid vaccine as a ... | bjp realli present free covid vaccin as a stat... | bjp realli pres free covid vac state ma |

## Sentiment Analysis

One assumption we made for this portion is that our sentiment analysis is treated to be the actual and correct sentiments from the tweets generated by VADER. All these sentiments will be used as the train & test inputs for our machine learning model for prediction.

In [5]:

```
##https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/##
# import SentimentIntensityAnalyzer class
# from vaderSentiment.vaderSentiment module.

import nltk
from nltk.sentiment.vader import SentimentIntensityAnalyzer
nltk.download('vader_lexicon')

# function to print sentiments
# of the sentence.
def sentiment_scores(sentence):

    # Create a SentimentIntensityAnalyzer object.
    sid_obj = SentimentIntensityAnalyzer()

    # polarity_scores method of SentimentIntensityAnalyzer
    # oject gives a sentiment dictionary.
    # which contains pos, neg, neu, and compound scores.
    sentiment_dict = sid_obj.polarity_scores(sentence)

#     print("Overall sentiment dictionary is : ", sentiment_dict)
#     print("sentence was rated as ", sentiment_dict['neg']*100, "% Negative")
#     print("sentence was rated as ", sentiment_dict['neu']*100, "% Neutral")
#     print("sentence was rated as ", sentiment_dict['pos']*100, "% Positive")

#     print("Sentence Overall Rated As", end = " ")

    # decide sentiment as positive, negative and neutral
    if sentiment_dict['compound'] >= 0.05 :
        return 1

    elif sentiment_dict['compound'] <= - 0.05 :
        return -1

    else :
        return 0

tweets_df['target'] = tweets_df['tweet_final'].apply(sentiment_scores)
```

```
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     C:\Users\gelat\AppData\Roaming\nltk_data...
[nltk_data]   Package vader_lexicon is already up-to-date!
```

In [6]:

```
tweets_df.head()
```

Out[6]:

| | date | username | tweet | tweet_cleaned | tweet_stem | tweet_fi |
|---|---|---|---|---|---|---|
| **0** | 2020-10-22 | to_fly_to_live | @ANI Isn't it the best poll promise ever?? Fre... | ani isnt it the best poll promise ever free co... | ani isnt it the best poll promis ever free cov... | ani isnt best promis ever f covid vac |
| **1** | 2020-10-22 | utkarshsinha07 | Now states shall have wait for thier Vidhan Sa... | now states shall have wait for thier vidhan sa... | now state shall have wait for thier vidhan sab... | state shall wait th vidhan sabha el ge |
| **2** | 2020-10-22 | batolebazi | जिस मदारी ने ट्रेन तक नहीं चलाई और तुम पत्नी व... | biharpolls covidvaccine httpstco5cshy52bvz | biharpol covidvaccin httpstco5cshy52bvz | bihar covidvac httpstco5cshy52 |
| **3** | 2020-10-22 | bak_sahil | @MisseeMonis They said vaccine for all but not... | misseemonis they said vaccine for all but not ... | misseemoni they said vaccin for all but not wh... | misseemoni s vaccin free cc vaccin new |
| **4** | 2020-10-22 | ivibhatweedy | BJP really presenting "free COVID vaccine" as ... | bjp really presenting free covid vaccine as a ... | bjp realli present free covid vaccin as a stat... | bjp realli pres free covid vac state ma |

## Dow Jones Index Average Dataset

The dataset includes the daily historical prices which spans across 2020 YTD.

Below is a list of the key terms within the dataset.

**1. Date** - The date from which data exists.

**2. Open** - The price of the Dow Jones Index at the start of the day's trading

**3. High** - The Highest Price reached within the particular date

**4. Low** - The Lowest Price reached within the particular date

**5. Close** - The price at the end of the day's trading

**6. Adj Close** - The price adjusted for changes in

**7. Volume** - The amount which changes across the course of a day.

We seek to obtain an additional statistic which can be used to measure changes in the stock price; the percentage change in a particular day. This, combined with the Date and the Volume, will give an idea of market sentiment involved.

In [76]:

```
dowjonesindex_df = pd.read_csv('^DJI.csv')

#Data cleaning for Dow Jones Index Average Dataset

def date_cleaned(date):
    date = pd.to_datetime(date, format='%Y-%m-%d')
    return date

dowjonesindex_df['Date'] = dowjonesindex_df['Date'].apply(date_cleaned)

dowjonesindex_df = dowjonesindex_df.rename(columns={'Date': 'date'})

# Percentage Price Change - Integer Encoding
# Positive = 1
# Negative = 0
# No change = 0

dowjonesindex_df['percentage_change'] = (dowjonesindex_df['Close'] - dowjonesindex_df[
'Open']) / dowjonesindex_df['Open'] * 100

dowjonesindex_df['price_label'] = dowjonesindex_df['percentage_change'].apply(lambda x:
1 if x > 0 else 0 if x < 0 else 0)

# Percentage Volume Change - Integer Encoding
# Positive = 1
# Negative = 0
# No change = 0

dowjonesindex_df['prev_Volume'] = dowjonesindex_df['Volume'].shift(1)

dowjonesindex_df['vol_percent_change'] = (dowjonesindex_df['Volume'] - dowjonesindex_df
['prev_Volume']) / dowjonesindex_df['prev_Volume'] * 100

dowjonesindex_df['volume_label'] = dowjonesindex_df['vol_percent_change'].apply(lambda
x: 1 if x > 0 else 0 if x < 0 else 0)

dowjonesindex_df = dowjonesindex_df.round(2)

dowjonesindex_df.head(10)
```

Out[76]:

| | date | Open | High | Low | Close | Adj Close | Volume | percentage_change |
|---|---|---|---|---|---|---|---|---|
| **0** | 2020-01-02 | 28638.97 | 28872.80 | 28627.77 | 28868.80 | 28868.80 | 251820000 | 0.80 |
| **1** | 2020-01-03 | 28553.33 | 28716.31 | 28500.36 | 28634.88 | 28634.88 | 239590000 | 0.29 |
| **2** | 2020-01-06 | 28465.50 | 28708.02 | 28418.63 | 28703.38 | 28703.38 | 252760000 | 0.84 |
| **3** | 2020-01-07 | 28639.18 | 28685.50 | 28565.28 | 28583.68 | 28583.68 | 258900000 | -0.19 |
| **4** | 2020-01-08 | 28556.14 | 28866.18 | 28522.51 | 28745.09 | 28745.09 | 291750000 | 0.66 |
| **5** | 2020-01-09 | 28851.97 | 28988.01 | 28844.31 | 28956.90 | 28956.90 | 275060000 | 0.36 |
| **6** | 2020-01-10 | 28977.52 | 29009.07 | 28789.10 | 28823.77 | 28823.77 | 237830000 | -0.53 |
| **7** | 2020-01-13 | 28869.01 | 28909.91 | 28819.43 | 28907.05 | 28907.05 | 249830000 | 0.13 |
| **8** | 2020-01-14 | 28895.50 | 29054.16 | 28872.27 | 28939.67 | 28939.67 | 287440000 | 0.15 |
| **9** | 2020-01-15 | 28901.80 | 29127.59 | 28897.35 | 29030.22 | 29030.22 | 260270000 | 0.44 |

## Data Visualisations

Before moving on to the methodology, visualisations were generated for the Dow Jones Index, Price Percentage Change and Volume from Jan 2020 to Dec.

In addition, we want to look at how the stock market reacted when news headlines related to Covid-19 Vaccines were published by several potential vaccine candidates during 2020.

Here are the timeline of the 3 main potential vaccine candidates which we selected that successfully developed a vaccine which has proven safe and effective by US Food & Drug Authority(FDA).

Timeline

**1. Moderna Therapeutics**

March 16: First patient dosed in a Phase 1 trial.

April 16: Moderna plans to begin a Phase 2 study by July.

May 1: Moderna announces it will partner with Swiss firm Lonza on development.

May 7: FDA clears Moderna to start a 600-patient Phase 2 study, which will begin "shortly".

May 18: Moderna discloses interim Phase 1 data, in which eight volunteers developed antibodies to the coronavirus.

May 29: Moderna doses the first volunteers in a Phase 2 study, planning to enroll about 600 people.

July 14: Moderna publishes Phase 1 data showing a consistent antibody response and mild to moderate side effects.

July 27: Moderna begins enrollment in a 30,000-subject Phase 3 trial.

Nov. 16: Moderna says its vaccine is 94.5% effective in the first look at data from its Phase 3 study.

Nov. 30: Moderna's vaccine showed 94% effectiveness in the Phase 3 trial's final efficacy analysis, and the company expects to submit for emergency authorization in the coming days.

**2. Pfizer & BioNTech**

April 9: BioNTech says it will begin its first human trials "as early as the end of April".

April 29: BioNTech and Pfizer dose the first patients in a Phase 1 trial in Germany, planning to enroll about 200 patients.

May 5: BioNTech and Pfizer begin Phase 1 study in the U.S., recruiting up to 360 patients total.

May 12: BioNTech says it expects preliminary data in June or July.

July 1: In a Phase 1 trial, BioNTech's vaccine led to an increase coronavirus antibodies at three doses, according to a preprint paper.

July 27: Pfizer and BioNTech begin enrollment on a 30,000-volunteer study, expecting data as early as October.

Sept. 12: Pfizer and BioNTech announce a plan to expand enrollment to 44,000 participants.

Nov. 9: Pfizer and BioNTech's vaccine was found to be more than 90% effective in the first analysis of Phase 3 data, the companies said.

Nov. 20: Pfizer and BioNTech submitted their vaccine for emergency use authorization in the U.S.

**3. AstraZeneca & University of Oxford**

March 27: Oxford begins recruiting patients for a placebo-controlled trial that will enroll up to 510 healthy volunteers. The vaccine will not be ready for "some weeks," according to the university.

April 30: AstraZeneca announces it will partner with Oxford to develop the vaccine.

July 20: Oxford publishes Phase 1/2 data demonstrating an immune response with mild to moderate side effects.

Aug. 31: AstraZeneca begins enrollment in a U.S. Phase 3 trial that will involved 30,000 volunteers.

Sept. 8: AstraZeneca says a hold has been put on the trial following a suspected adverse reaction in a participant.

Sept. 12: AstraZeneca says trials in the U.K. have resumed.

In [154]:

```
# Plot DJI Index and Volume Chart

# Use Date for indexing
dowjonesindex_df['date'] = pd.to_datetime(dowjonesindex_df.date,format="%Y-%m-%d")
dowjonesindex_df.index = dowjonesindex_df['date']

fig = plt.figure(figsize=(18, 15))

fig.add_subplot(3,1,1)
plt.plot(dowjonesindex_df['Close'], color='blue')
plt.grid(True)
plt.xlabel('date')
plt.ylabel('Index')
plt.title('Dow Jones Index')

plt.axvline("2020-07-14", 0, 1,
            label='14 July', color='black', linestyle='--')

plt.axvline("2020-11-16", 0, 1,
            label='16 Nov', color='black', linestyle='--')

plt.axvline("2020-07-01", 0, 1,
            label='01 July', color='purple', linestyle='--')

plt.axvline("2020-11-09", 0, 1,
            label='09 July', color='purple', linestyle='--')

plt.axvline("2020-07-20", 0, 1,
            label='20 July', color='brown', linestyle='--')

plt.axvline("2020-09-08", 0, 1,
            label='08 Sep', color='brown', linestyle='--')

plt.axvline("2020-09-23", 0, 1,
            label='23 Sep', color='brown', linestyle='--')

plt.legend(loc='upper left',prop={'size': 15})

ax2 = fig.add_subplot(3,1,2)
plt.plot(dowjonesindex_df['percentage_change'], color='red')
plt.grid(True)
plt.xlabel('date')
plt.ylabel('Percent Change')
plt.title('Price Percent Change - %')

plt.axvline("2020-07-14", 0, 1,
            label='14 July', color='black', linestyle='--')

plt.axvline("2020-11-16", 0, 1,
            label='16 Nov', color='black', linestyle='--')

plt.axvline("2020-07-01", 0, 1,
            label='01 July', color='purple', linestyle='--')

plt.axvline("2020-11-09", 0, 1,
            label='09 July', color='purple', linestyle='--')

plt.axvline("2020-07-20", 0, 1,
            label='20 July', color='brown', linestyle='--')
```

```
plt.axvline("2020-09-08", 0, 1,
            label='08 Sep', color='brown', linestyle='--')

plt.axvline("2020-09-23", 0, 1,
            label='23 Sep', color='brown', linestyle='--')

plt.legend(loc='upper left',prop={'size': 15})

ax2 = fig.add_subplot(3,1,3)
plt.plot(dowjonesindex_df['Volume']/1000000, color='green')
plt.grid(True)
plt.xlabel('date')
plt.ylabel('Volume - Millions')
plt.title('Volume')

plt.axvline("2020-07-14", 0, 1,
            label='14 July', color='black', linestyle='--')

plt.axvline("2020-11-16", 0, 1,
            label='16 Nov', color='black', linestyle='--')

plt.axvline("2020-07-01", 0, 1,
            label='01 July', color='purple', linestyle='--')

plt.axvline("2020-11-09", 0, 1,
            label='09 July', color='purple', linestyle='--')

plt.axvline("2020-07-20", 0, 1,
            label='20 July', color='brown', linestyle='--')

plt.axvline("2020-09-08", 0, 1,
            label='08 Sep', color='brown', linestyle='--')

plt.axvline("2020-09-23", 0, 1,
            label='23 Sep', color='brown', linestyle='--')

plt.legend(loc='upper left',prop={'size': 15})

plt.tight_layout()
```


**Main Events for Covid-19 Vaccine News:**

**1. Moderna Therapeutics**

July 14: Moderna publishes Phase 1 data showing a consistent antibody response and mild to moderate side effects.

Nov. 16: Moderna says its vaccine is 94.5% effective in the first look at data from its Phase 3 study.</p>

**2. Pfizer & BioNTech**

<p style="color:purple";>July 1: In a Phase 1 trial, BioNTech's vaccine led to an increase coronavirus antibodies at three doses, according to a preprint paper.</p>

<p style="color:purple";>Nov. 9: Pfizer and BioNTech's vaccine was found to be more than 90% effective in the first analysis of Phase 3 data, the companies said.</p>

**3. AstraZeneca & University of Oxford**

<p style="color:brown";>July 20: Oxford publishes Phase 1/2 data demonstrating an immune response with mild to moderate side effects.</p>

<p style="color:brown";>Sept. 8: AstraZeneca says a hold has been put on the trial following a suspected adverse reaction in a participant.</p>

<p style="color:brown";>Nov. 23: AstraZeneca said its vaccine was about 70% effective, on average, in the first look at Phase 3 data.</p>

## Key Findings

One key findings we identified was the unusual high trading volume and stock price volatility during the period when the Dow Jones Index crashed significantly between month of February and March.

Next, we will move on to combine two separate independent dataframes together for our machine learning model.

### Merging

In [77]:

```
sentiment_df = pd.merge(tweets_df, dowjonesindex_df, on = 'date')
sentiment_df
```

Out[77]:

| | date | username | tweet | tweet_cleaned | tweet_stem |
|---|---|---|---|---|---|
| **0** | 2020-10-22 | to_fly_to_live | @ANI Isn't it the best poll promise ever?? Fre... | ani isnt it the best poll promise ever free co... | ani isnt it the best poll promis ever free cov... |
| **1** | 2020-10-22 | utkarshsinha07 | Now states shall have wait for thier Vidhan Sa... | now states shall have wait for thier vidhan sa... | now state shall have wait for thier vidhan sab... |
| **2** | 2020-10-22 | batolebazi | जिस मदारी ने ट्रेन तक नहीं चलाई और तुम पत्नी व... | biharpolls covidvaccine httpstco5cshy52bvz | biharpol covidvaccin httpstco5cshy52bvz |
| **3** | 2020-10-22 | bak_sahil | @MisseeMonis They said vaccine for all but not... | misseemonis they said vaccine for all but not ... | misseemoni they said vaccin for all but not wh... |
| **4** | 2020-10-22 | ivibhatweedy | BJP really presenting "free COVID vaccine" as ... | bjp really presenting free covid vaccine as a ... | bjp realli present free covid vaccin as a stat... |
| **5** | 2020-10-22 | paulwatson72 | Another dose of daily miserablism from Planet ... | another dose of daily miserablism from planet ... | anoth dose of daili miserabl from planet graun... |
| **6** | 2020-10-22 | adarshshastri | The shame-facedness of the BJP has crossed all... | the shamefacedness of the bjp has crossed all ... | the shamefaced of the bjp ha cross all boundar... |
| **7** | 2020-10-22 | baskaranbharath | Just 2 days ago PM said roadmap is ready to pr... | just 2 days ago pm said roadmap is ready to pr... | just 2 day ago pm said roadmap is readi to pro... |
| **8** | 2020-10-22 | skphotography68 | #COVID vaccine. #vaccinepolitics #BiharElecti... | covid vaccine vaccinepolitics biharelections2020 | covid vaccin vaccinepolit biharelections2020 |
| **9** | 2020-10-22 | aamaadmiparty | What about non-BJP ruled states? Indians who... | what about nonbjp ruled states indians who d... | what about nonbjp rule state indian who didnt ... |
| **10** | 2020-10-22 | clivebennett | @theJeremyVine And on the same day we heard th... | thejeremyvine and on the same day we heard tha... | thejeremyvin and on the same day we heard that... |
| **11** | 2020-10-22 | akshay_tlkr | Next maharashtra assembly election will be hel... | next maharashtra assembly election will be hel... | next maharashtra assembl elect will be held in... |
| **12** | 2020-10-22 | paulhannon29 | Big Pharma, big money, big egos and a 900-year... | big pharma big money big egos and a 900yearold... | big pharma big money big ego and a 900yearold ... |
| **13** | 2020-10-22 | unusual_times | #Covid19UK #CovidVaccine Latest on the Astra... | covid19uk covidvaccine latest on the astraze... | covid19uk covidvaccin latest on the astrazenec... |
| **14** | 2020-10-22 | indtushar28 | Free Covid Vaccine for Bihar, With the kind of... | free covid vaccine for bihar with the kind of ... | free covid vaccin for bihar with the kind of c... |
| **15** | 2020-10-22 | cchandramouli1 | Misinterpretation. Misinformation. Beware #B... | misinterpretation misinformation beware biha... | misinterpret misinform bewar bihar polit parti... |
| **16** | 2020-10-22 | 30arihantparakh | Read @BJP4India 's manifesto for Bihar Electio... | read bjp4india s manifesto for bihar elections... | read bjp4india s manifesto for bihar elect cov... |

| | date | username | tweet | tweet_cleaned | tweet_stem |
|---|---|---|---|---|---|
| **17** | 2020-10-22 | tavysingh | Finance minister @nsitharaman launches manifes... | finance minister nsitharaman launches manifest... | financ minist nsitharaman launch manifesto for... |
| **18** | 2020-10-22 | tolkric | If you're pinning your hopes on a Covid vaccin... | if youre pinning your hopes on a covid vaccine... | if your pin your hope on a covid vaccin here a... |
| **19** | 2020-10-22 | dimacphail | If you're pinning your hopes on a Covid vaccin... | if youre pinning your hopes on a covid vaccine... | if your pin your hope on a covid vaccin here a... |
| **20** | 2020-10-22 | stup1dbarb1e | i wonder when there will be a covid vaccine | i wonder when there will be a covid vaccine | i wonder when there will be a covid vaccin |
| **21** | 2020-10-22 | raquelquefois | @jim_dickinson I've heard he died covid HOWEVE... | jimdickinson ive heard he died covid however h... | jimdickinson ive heard he die covid howev he h... |
| **22** | 2020-10-22 | hemagazineindia | @journoarunima @netshrink @doctorsoumya @c_ass... | journoarunima netshrink doctorsoumya cassisi t... | journoarunima netshrink doctorsoumya cassisi t... |
| **23** | 2020-10-22 | sufiyan_atif | @NairShilpa1308 @Ahmedshabbir20 And If BJP doe... | nairshilpa1308 ahmedshabbir20 and if bjp doesn... | nairshilpa1308 ahmedshabbir20 and if bjp doesn... |
| **24** | 2020-10-22 | imtiyaz94433919 | @ANI Covid Vaccine is not BJP's personal asset... | ani covid vaccine is not bjps personal asset | ani covid vaccin is not bjp person asset |
| **25** | 2020-10-22 | icehinvest_news | Oxford Developed Covid Vaccine, Then Scholars ... | oxford developed covid vaccine then scholars c... | oxford develop covid vaccin then scholar clash... |
| **26** | 2020-10-22 | salilshetty | Hope @yadavtejashwi will take #BJP to court an... | hope yadavtejashwi will take bjp to court and ... | hope yadavtejashwi will take bjp to court and ... |
| **27** | 2020-10-22 | jknewstoday | Covid vaccine trials can't tell if the shots s... | covid vaccine trials cant tell if the shots sa... | covid vaccin trial cant tell if the shot save ... |
| **28** | 2020-10-22 | sufiyan_atif | And If BJP doesn't win in Bihar will they char... | and if bjp doesnt win in bihar will they charg... | and if bjp doesnt win in bihar will they charg... |
| **29** | 2020-10-22 | ts_singhdeo | Free Covid vaccine is a right of every citizen... | free covid vaccine is a right of every citizen... | free covid vaccin is a right of everi citizen ... |
| **...** | ... | ... | ... | ... | ... |
| **161139** | 2020-02-27 | sherifazuhur | Fauci of the #NIH estimates a #COVID vaccine w... | fauci of the nih estimates a covid vaccine wil... | fauci of the nih estim a covid vaccin will lik... |
| **161140** | 2020-02-27 | soonergrunt | COVID-19 is 10 to 20 times more lethal than th... | covid19 is 10 to 20 times more lethal than the... | covid19 is 10 to 20 time more lethal than the ... |
| **161141** | 2020-02-27 | frencovfefe | COVID vaccine would take about about a year at... | covid vaccine would take about about a year at... | covid vaccin would take about about a year at ... |
| **161142** | 2020-02-27 | goshofar | @TheTNHoller @SecAzar If vaccine is not afford... | thetnholler secazar if vaccine is not affordab... | thetnhol secazar if vaccin is not afford the e... |

| | date | username | tweet | tweet_cleaned | tweet_stem |
|---|---|---|---|---|---|
| **161143** | 2020-02-27 | commentoniowa | @SenGillibrand @SenGillibrand, with plan let's... | sengillibrand sengillibrand with plan lets tra... | sengillibrand sengillibrand with plan let tran... |
| **161144** | 2020-02-27 | doctorworm3 | @whicks7667 @pablo_honey1 @realDonaldTrump @CN... | whicks7667 pablohoney1 realdonaldtrump cnn cdc... | whicks7667 pablohoney1 realdonaldtrump cnn cdc... |
| **161145** | 2020-02-27 | pudstah | @jeffereytweets @opusmarta @teemaloney @PaulGa... | jeffereytweets opusmarta teemaloney paulgaviga... | jeffereytweet opusmarta teemaloney paulgavigan... |
| **161146** | 2020-02-26 | suffjeff | Glass of wine gossip is that a covid vaccine i... | glass of wine gossip is that a covid vaccine i... | glass of wine gossip is that a covid vaccin is... |
| **161147** | 2020-02-26 | wizardbri | Bet there will be lines wrapper around buildin... | bet there will be lines wrapper around buildin... | bet there will be line wrapper around build to... |
| **161148** | 2020-02-26 | burberrywons | what do they expect her to do lmao it's not li... | what do they expect her to do lmao its not lik... | what do they expect her to do lmao it not like... |
| **161149** | 2020-02-26 | j_lemiech | @hap317 @Timcast IMO she should volunteer for ... | hap317 timcast imo she should volunteer for co... | hap317 timcast imo she should volunt for covid... |
| **161150** | 2020-02-26 | how_sustainable | COVID-19 Vaccine Shipped, and Drug Trials Star... | covid19 vaccine shipped and drug trials start ... | covid19 vaccin ship and drug trial start covid... |
| **161151** | 2020-02-26 | lymanstoneky | Again, a COVID vaccine candidate has already b... | again a covid vaccine candidate has already be... | again a covid vaccin candid ha alreadi been de... |
| **161152** | 2020-02-26 | kaavvz | Apparently an American company Moderna Therape... | apparently an american company moderna therape... | appar an american compani moderna therapeut a ... |
| **161153** | 2020-02-26 | lymanstoneky | I'm not watching the debate but I wanted to me... | im not watching the debate but i wanted to men... | im not watch the debat but i want to mention t... |
| **161154** | 2020-02-26 | cblevinems | @IHAVETHEANSWER3 @NIAIDNews Absolutely! Let me... | ihavetheanswer3 niaidnews absolutely let me ma... | ihavetheanswer3 niaidnew absolut let me make s... |
| **161155** | 2020-02-26 | lymanstoneky | @fordcharles Good thing there’s no COVID vacci... | fordcharles good thing theres no covid vaccine... | fordcharl good thing there no covid vaccin yet |
| **161156** | 2020-02-25 | doddsmaz | @ladyaimless1 @roseg What do they test the cov... | ladyaimless1 roseg what do they test the covid... | ladyaimless1 roseg what do they test the covid... |
| **161157** | 2020-02-25 | thinking_panda | I hope this vaccine works! It will bring hope ... | i hope this vaccine works it will bring hope t... | i hope thi vaccin work it will bring hope to a... |
| **161158** | 2020-02-25 | dowell_reuben | @lisaabramowicz1 Development of a Covid vaccin... | lisaabramowicz1 development of a covid vaccine... | lisaabramowicz1 develop of a covid vaccin is w... |
| **161159** | 2020-02-24 | page08 | #coronavirus #COVID #VACCINE Vaccine being tested | coronavirus covid vaccine vaccine being tested | coronaviru covid vaccin vaccin be test |

| | date | username | tweet | tweet_cleaned | tweet_stem |
|---|---|---|---|---|---|
| **161160** | 2020-02-24 | _rosebrush | sothis give me the patience not to argue with ... | sothis give me the patience not to argue with ... | sothi give me the patienc not to argu with the... |
| **161161** | 2020-02-24 | lalaruefrench75 | @VeritasDolor @ProAntiVaxxer @Relle36296199 No... | veritasdolor proantivaxxer relle36296199 noted... | veritasdolor proantivaxx relle36296199 note sc... |
| **161162** | 2020-02-14 | yelisa888 | They found a vaccine for the Coronavirus- and ... | they found a vaccine for the coronavirus and i... | they found a vaccin for the coronaviru and i f... |
| **161163** | 2020-02-14 | denisonlab | Short video on coronavirus vaccine development... | short video on coronavirus vaccine development... | short video on coronaviru vaccin develop effor... |
| **161164** | 2020-02-13 | derek_linders | @Laurie_Garrett Plenty of companies chasing he... | lauriegarrett plenty of companies chasing heat... | lauriegarrett plenti of compani chase heat mak... |
| **161165** | 2020-02-13 | allnbowtane | @girlsreallyrule Easy, researchers develop a C... | girlsreallyrule easy researchers develop a cov... | girlsreallyrul easi research develop a covid v... |
| **161166** | 2020-02-12 | leighnapier | @CaumontSimone @COVID_19NEWS @YouTube The SARS... | caumontsimone covid19news youtube the sars vac... | caumontsimon covid19new youtub the sar vaccin ... |
| **161167** | 2020-02-12 | p_anatacio | I guess the 18 months before the covid vaccine... | i guess the 18 months before the covid vaccine... | i guess the 18 month befor the covid vaccin is... |
| **161168** | 2020-02-12 | trudge1620 | @Grummz Yaya, now my insurance can get charged... | grummz yaya now my insurance can get charged f... | grummz yaya now my insur can get charg for a c... |

161169 rows × 18 columns

## III. Methodology - Exploratory Data Analysis & Machine Learning

In this section, we will explore the data in two parts.

Part 1: Using a Decision Tree Classification Machine Learning Model to compare the performance between price and volume changes in the Dow Jones Index.

Part 2: We will proceed to generate the heat map to derive a correlation followed by correlation equation for tweets sentiments and price & volume changes by using a linear regression model.

**Part 1: Decision Tree Classification Model**

In [78]:

```
sentiment_df.groupby('volume_label').size()
```

Out[78]:

```
volume_label
0    78267
1    82902
dtype: int64
```

In [79]:

```
sentiment_df.groupby('price_label').size()
```

Out[79]:

```
price_label
0    72083
1    89086
dtype: int64
```

In [80]:

```
sentiment_df.groupby('target').size()
```

Out[80]:

```
target
-1    38581
 0    65816
 1    56772
dtype: int64
```

**Train Test Split for Price Changes**

- **First** argument: sentiment_df[['Target']] - Sentiment Analysis by COVID-19 Tweets (integer encoded)
- **Second** argument: sentiment_df['price_label'] - Daily negative or positive price change in Dow Jones Index (integer encoded)
- `random_state` is 10
- `test_size` is 0.2

In [88]:

```
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(sentiment_df[['target']], sentiment
_df['price_label'], random_state = 10 , test_size = 0.2)

print(x_train.shape, x_test.shape, y_train.shape , y_test.shape)
```

```
(128935, 1) (32234, 1) (128935,) (32234,)
```

In [103]:

```
# Applying Decision Tree Classifier model

from sklearn.tree import DecisionTreeClassifier

sentiment_clf = DecisionTreeClassifier()

sentiment_clf.fit(x_train, y_train)

predicted_sentiment = sentiment_clf.predict(x_test)

from sklearn.metrics import f1_score, classification_report, accuracy_score

print('Accuracy of Price model:', accuracy_score(y_test, predicted_sentiment).round(4))
```

```
Accuracy of Price model: 0.5533
```

**Train Test Split for Volume Changes**

- **First** argument: sentiment_df[['Target']] - Sentiment Analysis by COVID-19 Tweets (integer encoded)
- **Second** argument: sentiment_df['volume_label'] - Daily negative or positive volume change in Dow Jones Index (integer encoded)
- `random_state` is 10
- `test_size` is 0.2

In [107]:

```
from sklearn.model_selection import train_test_split

X_train, X_test, Y_train, Y_test = train_test_split(sentiment_df[['target']], sentiment
_df['volume_label'], random_state = 10 , test_size = 0.2)

print(X_train.shape, X_test.shape, Y_train.shape , Y_test.shape)
```

```
(128935, 1) (32234, 1) (128935,) (32234,)
```

In [108]:

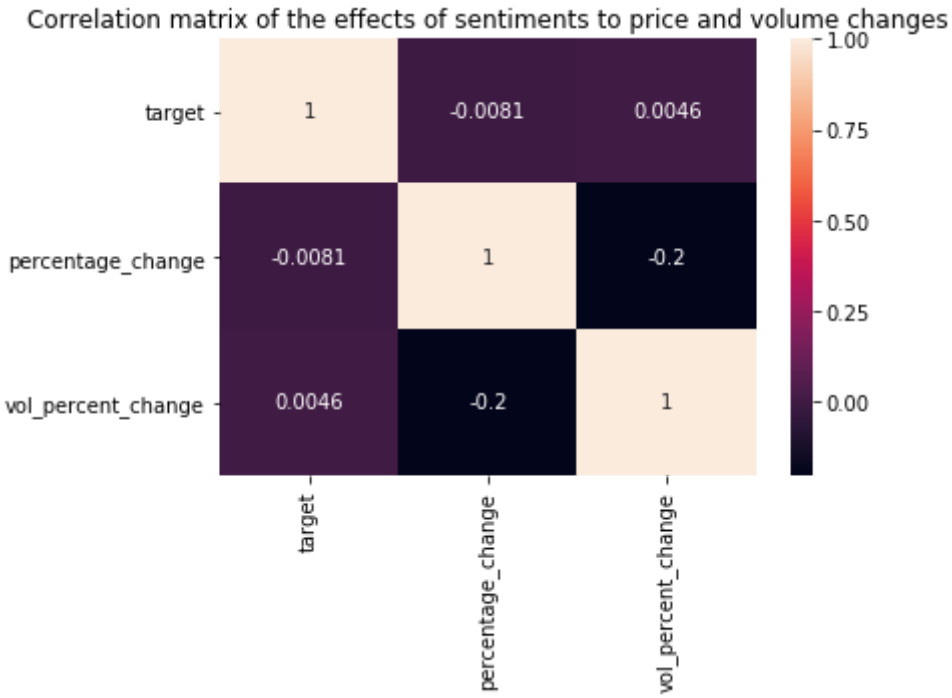
```
# Applying Decision Tree Classifier model

from sklearn.tree import DecisionTreeClassifier

sentiment_vol_clf = DecisionTreeClassifier()

sentiment_vol_clf.fit(X_train, Y_train)

predicted_vol_sentiment = sentiment_vol_clf.predict(X_test)

from sklearn.metrics import f1_score, classification_report, accuracy_score

print('Accuracy of Volume model:', accuracy_score(Y_test, predicted_vol_sentiment).roun
d(4))
```

```
Accuracy of Volume model: 0.5141
```

**Part 2A: Correlation Heat Map between Tweet Sentiment and Stock Market performance**

In [126]:

```
# Heat Map

# Target = Tweet Sentiments (Positive or Negative or Neutral)

# Percentage_change = Price Change (Positive or Negative)

# Vol_Percentage_change = Volume Change (Positive or Negative)

columns=['target','percentage_change','vol_percent_change']
correlations_df=sentiment_df[columns]
corrMatrix = correlations_df.corr()
sns.heatmap(corrMatrix, annot=True)

plt.title("Correlation matrix of the effects of sentiments to price and volume changes"
)

plt.show()
```




**Based on the heat map generated, the price change is more sensitive than volume change when there is sentiment change.**

**Part 2B: Linear Regression Equation - Price & Volume Change**

In [113]:

```
# Applying  Linear Regression model

from sklearn.linear_model import LinearRegression

sentiment_lr = LinearRegression()

sentiment_lr.fit(x_train, y_train)

print(sentiment_lr.coef_)
print(sentiment_lr.intercept_)

# Price change = 0.00656(tweet_sentiments) + 0.553

sentiment_vol_lr = LinearRegression()

sentiment_vol_lr.fit(X_train, Y_train)

print(sentiment_vol_lr.coef_)
print(sentiment_vol_lr.intercept_)

# Volume change = 0.00471(tweet_sentiments) + 0.514
```

```
[0.00065562]
0.5525303650918985
[0.00471888]
0.5139306034474946
```

## IV. Conclusion

Based on the Decision Tree Classifier Model, we can conclude that the model's performance is more accurate in predicting daily price changes compared to daily volume changes for Dow Jones Index.

**Accuracy**

1. Price Changes (%): 0.5533
2. Volume Changes (%): 0.5141

Based on the Linear Regression Model,

**1. Linear Regression Equation: Price Change (%) = 0.00656 * Tweet_sentiments + 0.553**

An increase in positive tweet sentiment (+1) will lead to a price increase of 0.0066%.

**2. Linear Regression Equation: Volume Change (%) = 0.00471 * Tweet_sentiments + 0.514**

An increase in positive tweet sentiment (+1) will lead to a volume increase of 0.0047%.

This justifies the correlation heat map generated earlier to conclude that price change is more sensitive to a volume change based on the percentages shown.

**Overall, a positive tweet relating to COVID-19 Vaccine News will lead to an uplift of the Dow Jones Index in both price and volume.**

## V. Further Analysis

After finding out the relationship between COVID-19 Vaccine News and Dow Jones Index, our team would like to dive deeper into researching:

1. The impact of COVID-19 Vaccine News on the various sectors to determine which sector is the least impacted and which sector is the most impacted.
2. The impact of COVID-19 on Dow Jones Index

## References

1. https://lionbridge.ai/datasets/top-10-stock-market-datasets-for-machine-learning/ (https://lionbridge.ai/datasets/top-10-stock-market-datasets-for-machine-learning/)
2. https://datainnovation.org/2020/08/tracking-the-development-of-covid-19-treatments-and-vaccines/ (https://datainnovation.org/2020/08/tracking-the-development-of-covid-19-treatments-and-vaccines/)
3. https://www.statnews.com/feature/coronavirus/drugs-vaccines-tracker/?utm_source=STAT+Newsletters&utm_campaign=4ae243c359-MR_COPY_01&utm_medium=email&utm_term=0_8cab1d7961-4ae243c359-149552929#vaccines (https://www.statnews.com/feature/coronavirus/drugs-vaccines-tracker/?utm_source=STAT+Newsletters&utm_campaign=4ae243c359-MR_COPY_01&utm_medium=email&utm_term=0_8cab1d7961-4ae243c359-149552929#vaccines)
4. https://www.kaggle.com/ritesh2000/covid19-vaccine-tweets (https://www.kaggle.com/ritesh2000/covid19-vaccine-tweets)
5. https://finance.yahoo.com/quote/%5EDJI/history?p=%5EDJI (https://finance.yahoo.com/quote/%5EDJI/history?p=%5EDJI)