Introduction

The pandemic of COVID-19 has affected the lives of people all over the globe, as well as disrupted economic activities and created difficult times. This initiative focuses on a particular COVID variant B 1.1.529 called Omicron. We are interested in using social media (Twitter) to acquire data regarding messages containing specific hashtags related to omicron. Here, we analyse the sentiments of Twitter users based on their use of hashtags.

Data Description and Source

Our dataset originates from Kaggle (URL: https://www.kaggle.com/gpreda/omicron-rising). This query collected approximately 17000 tweets containing the hashtags #Omicron, #COVID19, etc. via the Twitter API and a Python script. We have multiple fields in our data, including the user's name, location, tweets, and whether or not the user is verified. Using Sentiment and Regression Analysis, we analyse the data and provide answers to our research questions by combining all of these disciplines.

#	Variable	Description	Type
1	Id	Unique identifier for each tweet	Categorical
2	User_Name	Twitter profile name	Text
3	User_Location	Twitter profile location	Text
4	User_Description	Twitter profile bio	Text
5	User_Created	Date and time Twitter account was created	Date/Time
6	User_Followers	Number of Twitter followers	Numeric
7	User_Friends	Number of friends followed on Twitter	Numeric
8	User_Favorites	Number of tweets marked as a favorite	Numeric
9	User_Verified	Twitter account of public interest that is authentic	Binary
10	Date	Date and time tweet was posted	Date/Time
11	Text	Message posted to Twitter	Text
12	Hashtags	Keyword or phrase to group conversations on Twitter	Text
13	Source	Specific Twitter application used	Text
14	Retweets	Number of times the tweet was reposted	Numeric
15	Favorites	Number of times the tweet was marked as a favorite	Numeric
16	Is_Retweet	If this is a reposted tweet	Binary

Research Questions

This dataset will help to better understand how Twitter users reacted to the rise of Omicron. Specifically, we want to know:

- Are there more positive or negative reactions to Omicron?
- Are there any associations between positive and negative tweets about Omicron?
- Are there groups of Twitter users more likely to react positively? Or negatively?
- Are there are any geographic trends?
- Where are more users tweeting about Omicron from?
- Are there any day/time trends?
- When are more users tweeting about Omicron during the day?

Methodology

We will use two distinct algorithms to answer our research questions

- Sentiment Analysis
- Clustering

Sentiment Analysis is a natural language processing (NLP) technique that determines the positivity, neutrality, or negativity of text data. It is an essential business intelligence tool for gaining a deeper understanding of consumer trends and experiences. We intend to use Python code and Valence Aware Dictionary for sentiment Reasoning (VADER) to ascertain the aggregate sentiment expressed by Twitter users regarding the Omicron variant.

Clustering is an unsupervised machine learning algorithm that creates groups of data elements that are highly similar within a dataset. It is effective at organising data in order to construct meaningful structures, discover hidden patterns, and obtain deeper insights. We intend to use the Python library Scikit to apply clustering to the dataset in order to classify Twitter users according to whether they are more positive, neutral, or negative regarding Omicron.

Results and Description:

Overall Sentimental Analysis:

```
#Sentiment analysis of complete data

df1=review

df1['text'] = df1['text'].astype(str)

a='. '
for i in range(0,len(df1)):
    a=a+review['text'][i]
blob = TextBlob(a)
blob.sentiment
```

```
Sentiment(polarity=0.11308686786397787, subjectivity=0.4949662635754508)
```

Polarity is float, which falls within the range [-1,1], where 1 indicates a positive statement and -1 a negative one. Here, the polarity of sentiment is.1130, i.e., the data for the omicron variant of COVID-19 are neither overly positive nor negative, but incline, on average, towards a neutral evaluation.

Subjective clauses typically allude to personal opinion, emotion, or judgement, whereas objective clauses refer to factual data. Here, the Sensitivity analysis reveals that the evaluations are dominated by subjective opinions, feelings, or judgements.

The dataset has following distribution among the dataset:

- Positive review count: 5356
- Negative review count: 2075
- Neutral review count: 9615

Positive Topics and dataset distribution:

```
positive_topics #Headlines with positive sentiments
```

```
['covid give case would last high omicron always know straight promise take variant corona disease',
 'omicron covid vaccine people infection mask live india pandemic could better effective million still delta',
 'omicron variant coronavirus come latest mild keep update study look virus lead find today available',
 'case report omicron death covid today time booster news many update daily year even right',
 'omicron test positive good wave health covid evidence show strong rate first child week early']
```

```
df_doc_ptopic['dominant_topic'].value_counts()
```

```
2    1053
1     793
3     746
0     605
4     552
Name: dominant_topic, dtype: int64
```

Negative Topics and dataset distribution:

```
negetive_topics #Headlines with positive sentiments
```

```
['omicron covid previous long tell infection know news truth pandemic could fake medium time report',
 'record late vaccine lockdown track another early health trace didoing glib meal rishi unlock avoid',
 'omicron case coronavirus virus world report mean take bring around india cold sadly military death',
 'omicron case past death infect population worse covid hour mask thing serious total vaccination booster',
 'omicron variant come still base detrick fort year need test study truth people dangerous wave']
```

```
df_doc_ntopic['dominant_topic'].value_counts()
```

```
4    477
0    448
2    225
3    206
1     96
Name: dominant_topic, dtype: int64
```

Neutral Topics and dataset distribution:

```
neutral_topics #Headlines with positive sentiments
```

```
['omicron people pandemic fort detrick delta virus come still country government india minister take would',
 'covid insight analytics county team death case population confirm growth daily total distribution state life',
 'omicron need drericding report help cdcgov erictopol know mtosterholm covidwatch scottgottliebmd truth emergency time tell',
 'omicron vaccine variant covid show infection update study report news enough world restriction vaccination like',
 'omicron coronavirus give mask health case wave believe milder inevitable surrender surge variant today subvariant']
```

```
df_doc_neutopic['dominant_topic'].value_counts()
```

```
1    2061
3    1639
4    1255
0    1201
2     574
Name: dominant_topic, dtype: int64
```

Clustering analysis:

Text Data Analysis:

Initially creating 6 clusters:

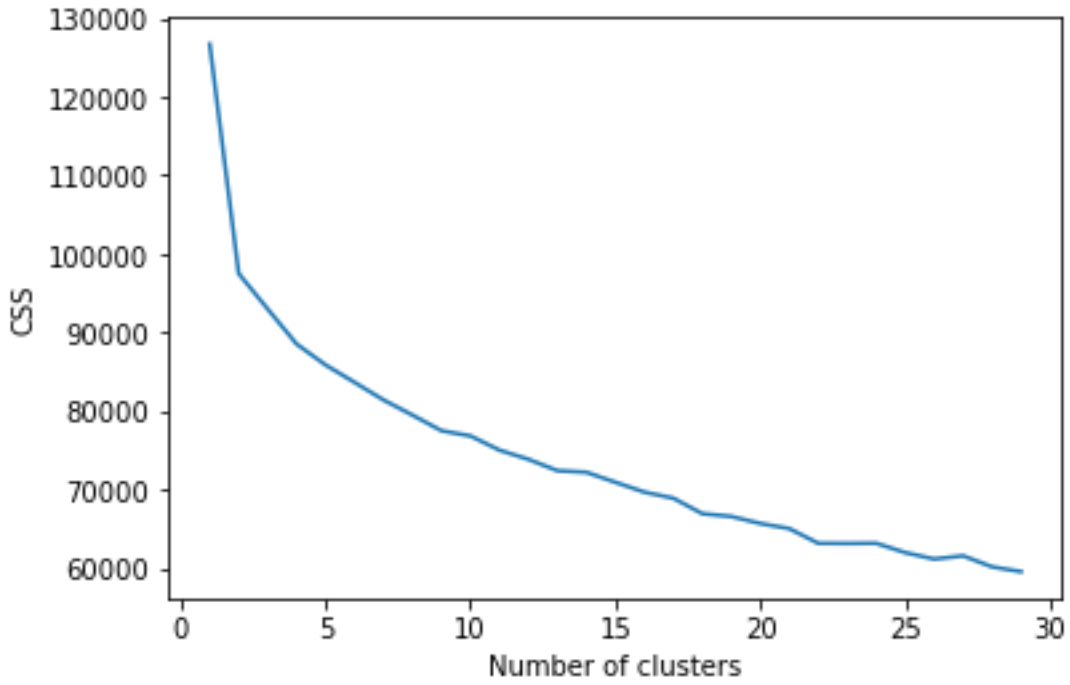
```
from sklearn.cluster import KMeans

NUM_CLUSTERS = 6
km = KMeans(n_clusters=NUM_CLUSTERS, max_iter=1000, n_init=50, random_state=42).fit(cv_matrix)
km
```

```
KMeans(max_iter=1000, n_clusters=6, n_init=50, random_state=42)
```

Clusters details:

```
CLUSTER #1
Key Features: ['new', 'case', 'new case', 'omicron', 'report', 'death', 'today', 'update', 'new new', 'covid', 'today new', 'da
ily', 'case death', 'case http', 'amp']
---------------------------------------------------------------------------
CLUSTER #2
Key Features: ['omicron', 'omicron http', 'covid', 'variant', 'omicron variant', 'case', 'get', 'vaccine', 'amp', 'new', 'covid
omicron', 'delta', 'coronavirus', 'wave', 'say']
---------------------------------------------------------------------------
CLUSTER #3
Key Features: ['covid', 'insight', 'covid insight', 'insight analytics', 'analytics', 'analytics team', 'team', 'county', 'team
http', 'county covid', 'death', 'covid death', 'population', 'distribution', 'day']
---------------------------------------------------------------------------
CLUSTER #4
Key Features: ['covid', 'case', 'amp', 'vaccine', 'new', 'come', 'death', 'fort', 'people', 'mask', 'detrick', 'fort detrick',
'say', 'test', 'report']
---------------------------------------------------------------------------
CLUSTER #5
Key Features: ['covid', 'county', 'per', 'covid insight', 'insight', 'population', 'per population', 'analytics', 'insight anal
ytics', 'county covid', 'population county', 'analytics http', 'day', 'case', 'growth']
---------------------------------------------------------------------------
CLUSTER #6
Key Features: ['covid', 'covid insight', 'insight', 'confirm', 'insight analytics', 'analytics', 'county', 'team', 'analytics t
eam', 'team http', 'county covid', 'confirm covid', 'case', 'case county', 'covid case']
---------------------------------------------------------------------------
```

Elbow Analysis:



After the analysis we can clearly say that 6 clusters are not sufficient and we can generate even 10 clusters to have a good set of clusters.

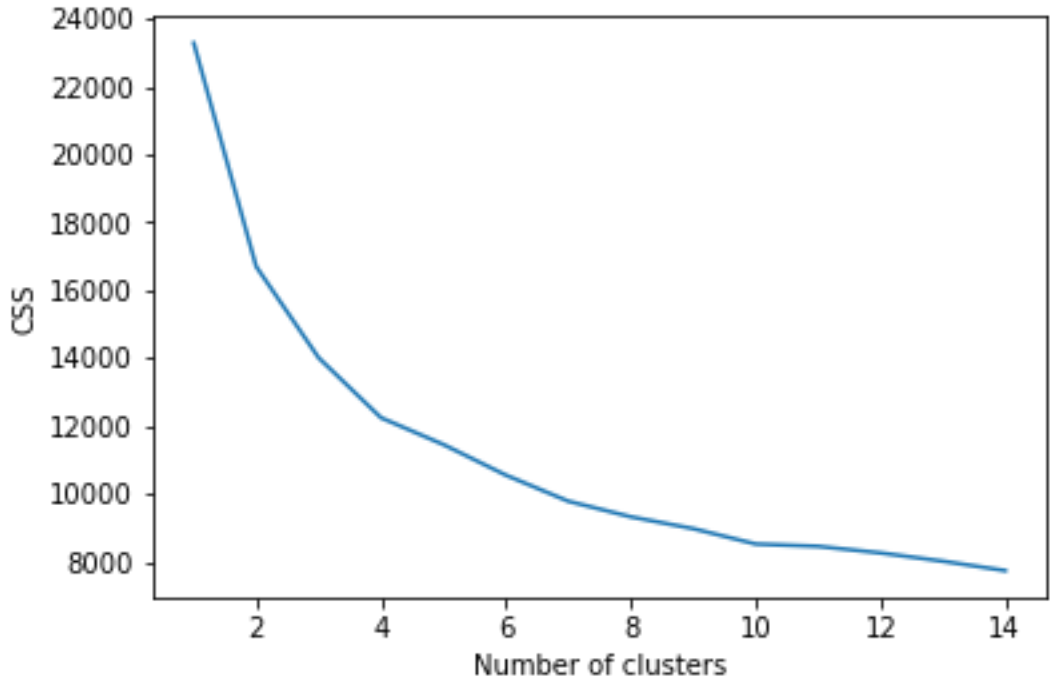
```
CLUSTER #1
Key Features: ['joyner', 'nathan joyner', 'nathan', 'erwin', 'explorer', 'express', 'feiglding', 'frank', 'frank west', 'frase
r']
---------------------------------------------------------------------------
CLUSTER #2
Key Features: ['dr', 'tomthunkit', 'times', 'md', 'health', 'business', 'india', 'jesus', 'gkay', 'gkay jesus']
---------------------------------------------------------------------------
CLUSTER #3
Key Features: ['save', 'save democracy', 'democracy', 'zerocovid', 'frank west', 'erwin loh', 'explorer', 'express', 'feigldin
g', 'frank']
---------------------------------------------------------------------------
CLUSTER #4
Key Features: ['join twtter', 'last', 'girly', 'swots', 'swots join', 'last girly', 'twtter', 'girly swots', 'join', 'displays
llc']
---------------------------------------------------------------------------
CLUSTER #5
Key Features: ['news', 'medical', 'news medical', 'english', 'thailand', 'world', 'mumbai', 'alert', 'china', 'global']
---------------------------------------------------------------------------
CLUSTER #6
Key Features: ['druider', 'bron druider', 'bron', 'zerocovid', 'erwin loh', 'express', 'feiglding', 'frank', 'frank west', 'fra
ser']
---------------------------------------------------------------------------
CLUSTER #7
Key Features: ['meadows', 'lisa marie', 'lisa', 'marie meadows', 'marie', 'fraser', 'explorer', 'express', 'feiglding', 'fran
k']
---------------------------------------------------------------------------
CLUSTER #8
Key Features: ['ann', 'ann marie', 'marie', 'pincivero', 'marie pincivero', 'fraser', 'express', 'feiglding', 'frank', 'frank w
est']
---------------------------------------------------------------------------
CLUSTER #9
Key Features: ['newsonline', 'zerocovid', 'eric feiglding', 'erwin loh', 'explorer', 'express', 'feiglding', 'frank', 'frank we
st', 'fraser']
---------------------------------------------------------------------------
CLUSTER #10
Key Features: ['raj', 'raj rajnarayanan', 'rajnarayanan', 'zerocovid', 'frank', 'eric feiglding', 'erwin', 'erwin loh', 'explor
er', 'express']
---------------------------------------------------------------------------
```

Hashtag Data Analysis:

Initially Creating 6 Clusters and their details:

```
CLUSTER #1
Key Features: ['covid19', 'covid', 'coronavirus', 'vaccine', 'hongkong', 'unitedstates', 'ba2', 'pandemic', 'corona', 'vaccinat
ion']
---------------------------------------------------------------------------
CLUSTER #2
Key Features: ['covid19', 'omicron', 'covid19 omicron', 'omicron covid19', 'coronavirus', 'covid', 'covid covid19', 'chorona',
'omicron chorona', 'covid19 covid19']
---------------------------------------------------------------------------
CLUSTER #3
Key Features: ['nan', 'zerocovid', 'covidvaccine', 'covidisairborne', 'covid19ab', 'covid19 variantdashboard', 'covid19 vaccin
e', 'covid19 unitedstates', 'covid19 pandemic', 'covid19 omicron']
---------------------------------------------------------------------------
CLUSTER #4
Key Features: ['omicron', 'covid', 'ba2', 'delta', 'coronavirus', 'omicron covid', 'omicron delta', 'omicron ba2', 'sarscov2',
'delta omicron']
---------------------------------------------------------------------------
CLUSTER #5
Key Features: ['phuket', 'thailand', 'thailand bangkok', 'bangkok', 'bangkok phuket', 'samui', 'phuket samui', 'samui pattaya',
'pattaya', 'ayutthaya']
---------------------------------------------------------------------------
CLUSTER #6
Key Features: ['covid', 'omicron', 'covid omicron', 'covid19', 'covid19 covid', 'covid covid', 'covid19 covid19', 'coronaviru
s', 'vimeo', 'omicron covid19']
---------------------------------------------------------------------------
```

Elbow Analysis:



The analysis of the above graph shows that we should have 7 clusters will be optimal for the proper cluster definition.

```
CLUSTER #1
Key Features: ['joyner', 'nathan joyner', 'nathan', 'erwin', 'explorer', 'express', 'feiglding', 'frank', 'frank west', 'frase
r']
---------------------------------------------------------------------------
CLUSTER #2
Key Features: ['news', 'dr', 'newsonline', 'druider', 'bron', 'bron druider', 'times', 'md', 'health', 'india']
---------------------------------------------------------------------------
CLUSTER #3
Key Features: ['meadows', 'lisa marie', 'lisa', 'marie meadows', 'marie', 'fraser', 'explorer', 'express', 'feiglding', 'fran
k']
---------------------------------------------------------------------------
CLUSTER #4
Key Features: ['ann', 'ann marie', 'marie', 'pincivero', 'marie pincivero', 'fraser', 'express', 'feiglding', 'frank', 'frank w
est']
---------------------------------------------------------------------------
CLUSTER #5
Key Features: ['news', 'medical', 'news medical', 'thailand', 'network', 'feiglding', 'erwin', 'erwin loh', 'explorer', 'expres
s']
---------------------------------------------------------------------------
CLUSTER #6
Key Features: ['tomthunkit', 'eric feiglding', 'erwin', 'erwin loh', 'explorer', 'express', 'feiglding', 'frank', 'frank west',
'zerocovid']
---------------------------------------------------------------------------
CLUSTER #7
Key Features: ['save', 'save democracy', 'democracy', 'zerocovid', 'frank west', 'erwin loh', 'explorer', 'express', 'feigldin
g', 'frank']
---------------------------------------------------------------------------
```

Geographical Analysis:

In this section we will be working on the geographical analysis and to see how people are reacting in the specific areas.

Tweet distribution:

```
1 loc_rev['user_location'].value_counts()
```

```
Los Angeles, CA                2658
India                           474
USA                             325
In Your Mind Now                211
Chandigarh                      210
                               ...
North Kingstown, RI               1
Global                            1
Kiev                              1
Pompano Beach, FL                 1
Auckland Region, New Zealand      1
Name: user_location, Length: 2474, dtype: int64
```

From the distribution its clear that most people tweeting are from Los Angeles, CA and is nearly 5X compared to that of India.

Also its visible that there is no specific format for location making it difficult to give any exact insight out of the data.

Analysis for Los Angeles, CA:

Polarity=0.3805, depicting that people have a positive response towards omicron compared to previous COVID-19 strains.

Subjectivity=0.9384, the dataset is highly subjective

Compared to the sentiment analysis of overall dataset LA people are more positive and show a trait of high openness to speak about personal thoughts.

Various Topics that have a dominance in LA region

```
1 LA_topics
```

```
['deaths 2022-02-10 population usa… growth tea… cases usafacts… confirmed 2022-02-14 daily state distribution total death',
 'distribution state 2022-02-14 total population 2022-02-08 death usafacts… usafac… deaths usa… daily 2022-02-10 growth cases',
 '2022-02-14 population growth cases tea… confirmed deaths daily usa… usafacts… 2022-02-10 distribution state death usafac…',
 'usafacts… 2022-02-14 cases confirmed daily deaths total 2022-02-08 growth distribution population 2022-02-10 tea… death usafa
c…',
 'confirmed growth cases 2022-02-10 2022-02-14 deaths population tea… state distribution death daily 2022-02-08 usafacts… tota
l']
```

Topic distribution (test data):

```
1 df_doc_LaTopic['dominant_topic'].value_counts()
```

```
2    522
3    476
1    329
0    279
4    254
Name: dominant_topic, dtype: int64
```

User Analysis:

Distribution of user data is as follows

```
1 user_rev['user_name'].value_counts()
```

```
Nathan Joyner                  2632
save DEMOCRACY                  282
Tomthunkit™                     211
Newsonline                      205
bron druider                    156
                               ...
Time Of India                     1
RestaurantOwner                   1
Pandemic-Aid Networks             1
Lieutenant General Ron Place      1
Kuldip Patel                      1
Name: user_name, Length: 6012, dtype: int64
```

Most active user on the taken set of tweeters is Nathan Joyner also he is from Los Angeles, CA so choosing him for analysis will provide us with same results as that of that. So, choosing the user “save DEMOCRACY” for analysis.

polarity=0.0703, subjectivity=0.5267, he is more or a neutral user have subjective tweets but is neither negative nor positive towards the omicron variant of COVID-19

```
1 NJ_topics
```

```
['that inevitable surrender give milder just omicron will believe fighti… cases promised enough indicator going',
 'what cases that would always covid straight going give promised conti… said… cases/capita keep indicator',
 'enough omicron that believe will indicator what come going leading inevitable would cases said… milder']
```

```
1 df_doc_NJTopic['dominant_topic'].value_counts()
```

```
0    87
1    69
2    41
Name: dominant_topic, dtype: int64
```

Conclusion:

In conclusion, our analysis of Twitter data regarding the Omicron variant of COVID-19 revealed that the majority of user sentiments were neutral, with an equal number of positive and negative reactions. Subjective opinions dominated the dataset, indicating that emotions played a significant role in Omicron-related discussions. Clustering assisted in classifying users based on their sentiments, thereby providing valuable insight into distinct user segments.

According to a geographical analysis, Los Angeles, California is the most engaged region with a favourable opinion of Omicron. However, it is essential to note that Twitter data may not reflect the views of the entire population.

This study contributes to policymakers' and health authorities' comprehension of public perceptions during a pandemic. By employing Sentiment Analysis and Clustering, we acquired a deeper understanding of how social media users reacted to Omicron. While the dataset contains valuable information, additional research utilising diverse data sources would yield a more complete comprehension. This study highlights the significance of monitoring social media sentiments during significant events in order to effectively comprehend public reactions and attitudes.