ASSIGNMENT COVERSHEET

UTS: ENGINEERING & INFORMATION TECHNOLOGY

SUBJECT NUMBER & NAME 41030 Engineering Capstone	**NAME OF STUDENT(s)** (PRINT CLEARLY) Koshin Jama	**STUDENT ID(s)** 12889426

STUDENT EMAIL koshin.jama@student.uts.edu.au	**STUDENT CONTACT NUMBER**

NAME OF TUTOR Wei Liu	**TUTORIAL GROUP** N/A	**DUE DATE** November 8th 2021

ASSESSMENT ITEM NUMBER & TITLE

Capstone Report

☐ I confirm that I have read, understood and followed the guidelines for assignment submission and presentation on page 2 of this cover sheet.
☐ I confirm that I have read, understood and followed the advice in the Subject Outline about assessment requirements.
☐ I understand that if this assignment is submitted after the due date it may incur a penalty for lateness unless I have previously had an extension of time approved and have attached the written confirmation of this extension.

Declaration of originality: The work contained in this assignment, other than that specifically attributed to another source, is that of the author(s) and has not been previously submitted for assessment. I understand that, should this declaration be found to be false, disciplinary action could be taken and penalties imposed in accordance with University policy and rules. In the statement below, I have indicated the extent to which I have collaborated with others, whom I have named.

Statement of collaboration:

Signature of student(s) Koshin Jama **Date** 31/10/2021

Table of Contents

1. Introduction: Engineering Problem

The usage of social media during global crises and natural disasters surges drastically; evidently seen through the early onset of the COVID-19 pandemic. Social media has become a crucial communication tool that has been used for information generation, consumption and expeditious information transmission. Amidst the height of the global COVID-19 pandemic in 2020, user's on 'twitter' took their ideals to publish their own 'tweets' on the platform that often clashed against other well established authoritative health figures in an attempt to combat the high influx of engagement created. In doing so, The World Health Organisation (WHO) has classified social media platforms to cause the first ever recorded phenomenon known as an 'infodemic' which is the occurrence of an overabundance of both accurate and inaccurate information being readily available. Individuals with access to 'high quality' information that are backed by internationally renowned institutions approved by governments are conflicted with spreading knowledge as it will further the infodemic. Thus in turn making it harder for user's who primarily rely on the strenuous use of social media platforms to acquire reliable information in a sea of misinformation being a catalyst in furthering the global pandemic.

The occurrence of the infodemic can be used to supplement our rather limited understanding of the ramifications that the pandemic had influenced on social media platforms such as twitter. By understanding the nature of tweets in relevance with COVID-19 we can identify patterns and trends in the behaviour of user's when 'tweeting'. Tweets naturally contain a variety of meta-data that can be extracted if a successful authorisation to the application programming interface (API) server can be established. This data can be mined intuitively in high volumes within a suitable data frame through the use of 'Natural Language Processing' (NLP) libraries that are pre-processed to allow visualisations to be designed. The illustrations made from the meta-data extracted from tweets can allow for comprehensible conclusions to be drawn when using licensed proprietary software to help exploit hidden relationships.

The project is to conceptualise and analyse tweets that are of relevance to COVID-19 using an existing open sourced twitter API server known as 'Tweepy'; to utilise the in-built functions to search for specific embedded hashtags. Select traits embedded within the tweet's meta-data will be saved to a data frame. Through the extensive use of NLP libraries such as 'pandas', the data will be filtered to account for a variety of false positives that are obtained unintentionally when mining large volumes of tweets. The data will be further processed to remove any noise to smooth out the data such as 'bots' polluting specific hashtags that are being investigated in the data set to be exported locally to a file. The file is then imported to a visualisation software to draw multiple graphs spanning between basic line graphs to understand the gradual progression of a particular hashtag performance throughout the year and more advanced visualisation techniques such as a stacked area chart.

The hashtags that are carefully studied throughout the duration of this project are: '#Lockdown, #Vaccine, #Mortality, #JobKeeper & #WFH'. Tweepy's limitation only allows tweets within 7 days to be extracted making it impossible to access data from the previous year when the 'infodemic' was at its all time high. Through careful planning and deliberation, multiple other trends were observed through analytical tools such as google adwords as these hashtags fell relatively short in having any sort of long term significance to be studied. These trends that followed this nature often fell off just as fast as they started making them redundant in being observed as the final listed trends provided worthwhile patterns that were observed.

2. Literature Review

Throughout the years, technology has become more prevalent in our daily lives as it becomes widely accessible worldwide. In its extensive use, the amount of meta-data generated within these electronic devices has drastically increased resulting in enormous data set's that lie dormant within databases. If the data has been processed or contains inconsistent redundant values that cannot be interpreted, every value within the data set contains a story that is yet to be interpreted. Whether the data set is generated through mining or extracted from a reputable source, the foremost method in capturing perception is through visualisation. "Data Visualisation Principles and Practice" by Alexandru C. Telea explores the early infancy stages of data visualisation to the mainstream adoption by business to make data driven decisions.

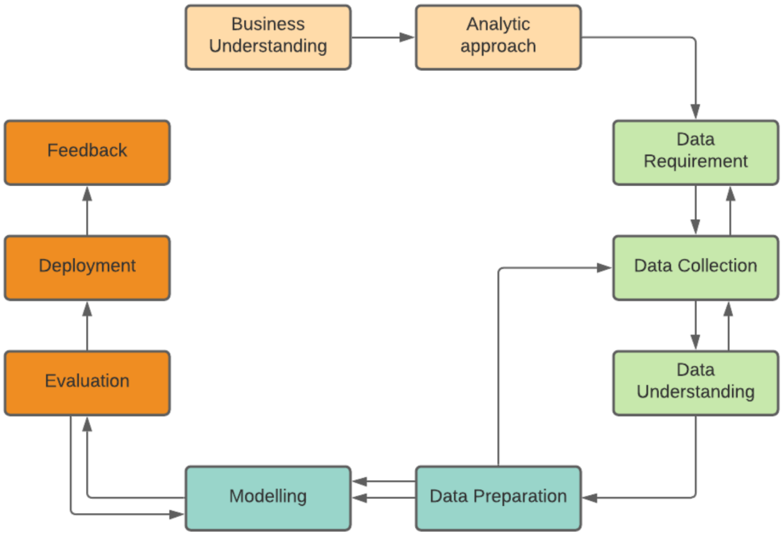
Alexandru eloquently describes the purpose of visualisation is to gather insight through qualitative measures through the use of interactive graphics. In the process of generating visual mediums, there are few questions that need to be addressed first before progressing. "Discovering the unknown" by being able to understand the nature of the dataset that is being sampled. Constantly prompting yourself by trying to find concrete answers about the problem being investigated is the first step taken to deepen your understanding. Moreover the other counterpart to the question is to quantitatively find "the data's minimum distribution" and qualitatively 'find answers in the unknown'. In reference to the nature of this capstone project, discovering the unknown is to be able to gain more depth on the nature of the dataset being created. This is a crucial step as the data set being mined from the tweepy API is not concrete as there is not enough data present to generate a final data set but allows for specific traits to be observed in judging its significance . Resulting in creating a rough architecture in which values will be taken into consideration to be used for further interpretations that can be drawn within it. Furthermore, to quantitatively measure the minimum distribution is to manipulate the final data to observe total engagement within the platform to gauge performance of relevant hashtags collectively. Finally, the data will qualitatively answer a majority of the unknowns within the data set as the visualisation made will expose hidden relationships between hashtags throughout the pandemic.

Subsequently upon the generation of the visualisation graphics, Alexandru prompts the reader to understand the difference in results from observing data taken quantitatively and qualitatively. There are certain instances where one method will provide a more clear insight in the data set than the other and it is crucial in understanding why this is the case. Gaining this understanding will help elevate the quality of reassurance in being able to make better data driven decisions that have started to become adopted mainstream within scientists and businesses alike. However, in most cases when measuring quantitative datasets they are often best matched with qualitative visualisations to provide a smoother interpretation as outliers within the data set can be studied to see if there are anomalies. In this project there will be little quantitative measures taken as most of the understanding will be derived from visualisations. Due to the nature of the dataset, quantitative assessments will not enhance the insight being drawn within the data set as there are not many worthwhile conclusions to be drawn from it. Multiple visualisation techniques will be used to complement this as it will allow for a more complex relationship to be seen throughout the growth of all five hashtags being studied throughout the year. Quantitatively analysing the dataset will not allow you to effectively study the relationship of the hashtags within one another.

3. Methodology

The process methodology that is to be adapted in this capstone project is the “Data Science Methodology”. This is a cyclic process that explores multiple solutions for a specific problem. The methodology begins with understanding the problem from an analytical perspective and understanding what data needs to be used for the problem. The data is then to be cleansed and prepared before it is exported. The solutions are subjective to the issue since they are drawn from the visualisation process, however it is important for discussion to occur amongst peers during this stage to suggest the best possible solution. This same methodology will be integrated within the project and all results drawn from the data obtained will be discussed carefully with the supervisor. Furthermore the nature of this project is split into three major components that are completed sequentially before moving which are the ‘Collection of Tweets’, ‘Preprocessing’ and finally ‘Visualisation & Analysis’.

Down below illustrates the data science methodology that was followed through to a certain extent:

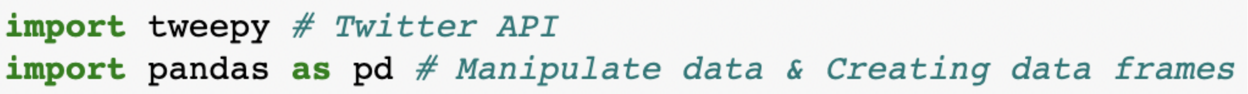


1. Collection of Tweets

A data pipeline needed to be created and established to successfully extract data from the Tweepy API in order to create a data set that can be visualised. This is the most crucial step as Tweepy does not allow tweets more than seven days old to be extracted meaning that the pipeline needs to be as efficient to reduce run time errors that could occur when being run weekly. To help mitigate this issue, jupyter notebook was the most appropriate tool in this scenario as it allows for multiple blocks of code to be modularised to execute certain functions for the required output needed to generate a data set. The python library Pandas was utilised to help create succinct data frames to host all the data being extracted.

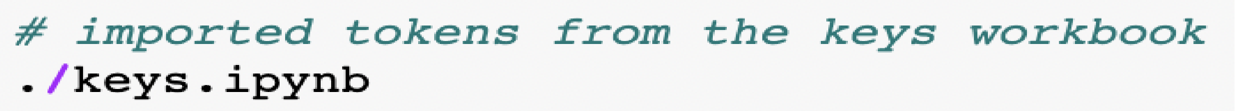
Excerpts of code that provide most of the functionality of this step needed to mine data and store it is extracted below:

Cell Block 1: Import statements that were initialised through the terminal to access the twitter servers and also manipulate data through pandas (referenced as pd throughout)



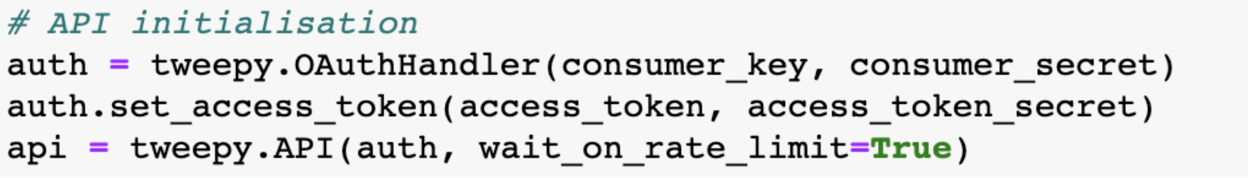
```
import tweepy # Twitter API
import pandas as pd # Manipulate data & Creating data frames
```

Cell Block 2: An individual jupyter notebook session that hosts a consumer_key, consumer_secret, access_token and access_token_secret variables that are imported to this main current workbook. These tokens are private and are only generated once throughout each project when registering for a developer account on twitter. For security reasons these will remain anonymous.



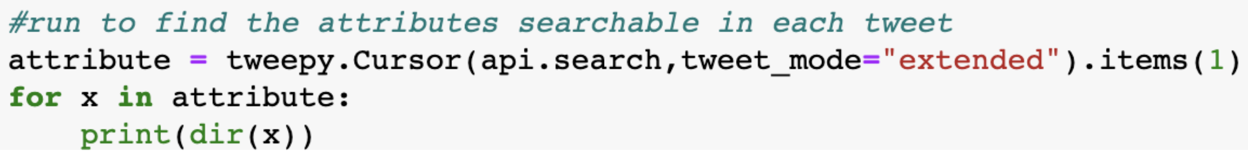
```
# imported tokens from the keys workbook
./keys.ipynb
```

Cell Block 3: Before the data can be accessed from the server, identification is needed to create authorisation access to it. To establish this we set up a couple of variables in this cell to host some tokens that were imported from the cell above and create the API variable that is used to make these requests.



```
# API initialisation
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth, wait_on_rate_limit=True)
```

Cell Block 4: To figure out which attributes give each tweet the meta-data that is to be extracted a simple cursor was created. A cursor is what is used to traverse all incoming tweets based on the parameters it takes in. In this instance, the latest tweet was taken since no parameters were established and observed to figure out how to access these specific elements in the list masked by the cursor.



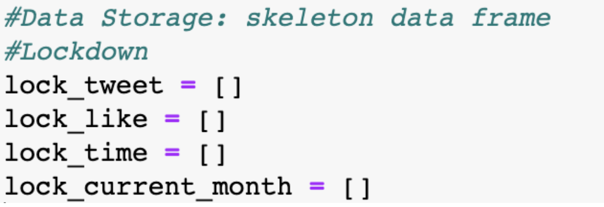
```
#run to find the attributes searchable in each tweet
attribute = tweepy.Cursor(api.search,tweet_mode="extended").items(1)
for x in attribute:
    print(dir(x))
```

The ‘dir()’ function is the reference to the directory that hosts all the meta-data attributes given to each tweet. A simple for loop was created to traverse all possible data that could be used within the final data set. The output of this executed cell:

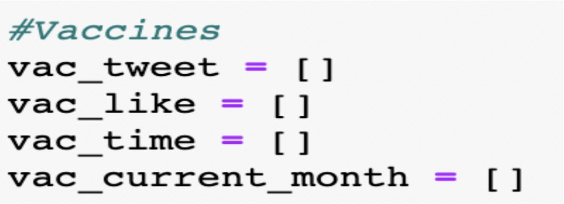
```
['__class__', '__delattr__', '__dict__', '__dir__', '__doc__', '__eq__', '__format__', '__ge__', '__getattribute__',
'__getstate__', '__gt__', '__hash__', '__init__', '__init_subclass__', '__le__', '__lt__', '__module__', '__ne__', '_
_new__', '__reduce__', '__reduce_ex__', '__repr__', '__setattr__', '__sizeof__', '__str__', '__subclasshook__', '__we
akref__', '_api', '_json', 'author', 'contributors', 'coordinates', 'created_at', 'destroy', 'display_text_range', 'e
ntities', 'favorite', 'favorite_count', 'favorited', 'full_text', 'geo', 'id', 'id_str', 'in_reply_to_screen_name',
'in_reply_to_status_id', 'in_reply_to_status_id_str', 'in_reply_to_user_id', 'in_reply_to_user_id_str', 'is_quote_sta
tus', 'lang', 'metadata', 'parse', 'parse_list', 'place', 'retweet', 'retweet_count', 'retweeted', 'retweeted_statu
s', 'retweets', 'source', 'source_url', 'truncated', 'user']
```

Upon first inspection, a lot of these attributes were attractive at first glance, however the majority of these variables hosted null values. Variables in here such as coordinates, and geo location would have been ideal to have in the final dataset to create a geographic heat map showing tweet activity however when extracting more test tweets; these variables all returned null values making them redundant.

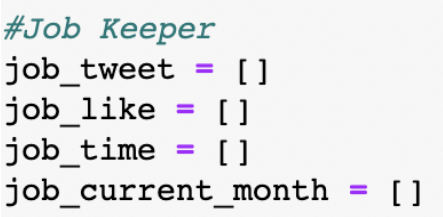
Cell Block 5: A skeleton data frame of variables was created to host the extracted data. These variables are to contain different values monthly as the uptime of the workbook was only restarted once a month once the variables were exported to a csv file.



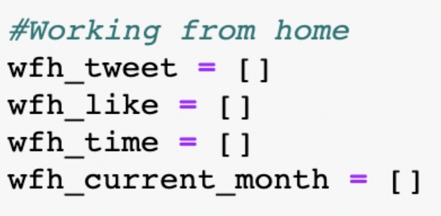
```
#Data Storage: skeleton data frame
#Lockdown
lock_tweet = []
lock_like = []
lock_time = []
lock_current_month = []
```



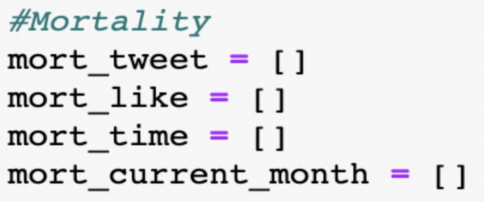
```
#Vaccines
vac_tweet = []
vac_like = []
vac_time = []
vac_current_month = []
```



```
#Job Keeper
job_tweet = []
job_like = []
job_time = []
job_current_month = []
```



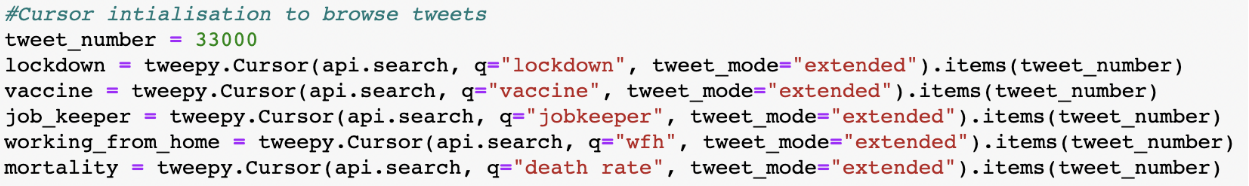
```
#Working from home
wfh_tweet = []
wfh_like = []
wfh_time = []
wfh_current_month = []
```



```
#Mortality
mort_tweet = []
mort_like = []
mort_time = []
mort_current_month = []
```

The variables that are taken from the directory were the tweet's text, likes and time it was published on the platform. These were the only fields that are required for a tweet to be sent and will always have a value that can be analysed.

Cell Block 6: A cursor was created for each hashtag that traversed the data being requested to the API. Each cursor had parameters filled in to search for relevant hashtags related to COVID-19 in general. All the cursors were added into a single cell for clarity and to reduce overall runtime. If an individual cursor needed to run individually the rest of the cursors are commented out.

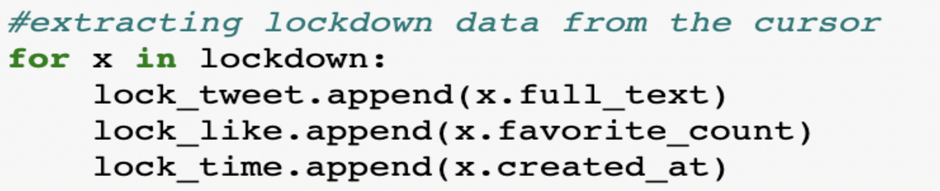


```
#Cursor intialisation to browse tweets
tweet_number = 33000
lockdown = tweepy.Cursor(api.search, q="lockdown", tweet_mode="extended").items(tweet_number)
vaccine = tweepy.Cursor(api.search, q="vaccine", tweet_mode="extended").items(tweet_number)
job_keeper = tweepy.Cursor(api.search, q="jobkeeper", tweet_mode="extended").items(tweet_number)
working_from_home = tweepy.Cursor(api.search, q="wfh", tweet_mode="extended").items(tweet_number)
mortality = tweepy.Cursor(api.search, q="death rate", tweet_mode="extended").items(tweet_number)
```

The variable tweet_number was created to quickly modify how many tweets will be requested and extracted from the API as one of the limitations of the basic twitter developer account caps monthly tweets at 500,000 per month.

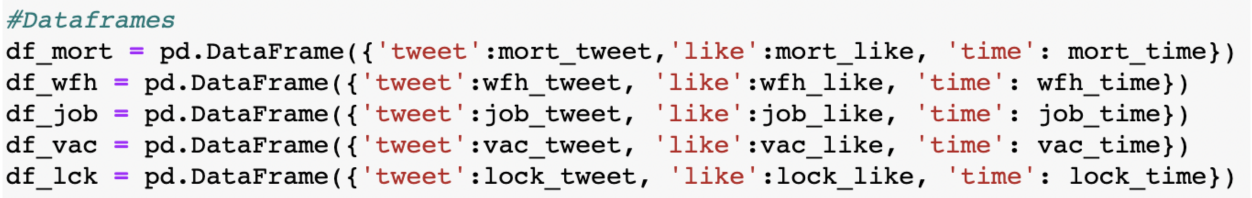
2. Preprocessing

Cell Block 7: Once all the pipeline measures are correctly implemented above, the dataset needs to be processed. To start cleaning the data we need to store the data of all the variables in the skeleton framework in order to develop the dataframe to be pushed onto a csv file. This same block of code was repeated for every other hashtag being studied.



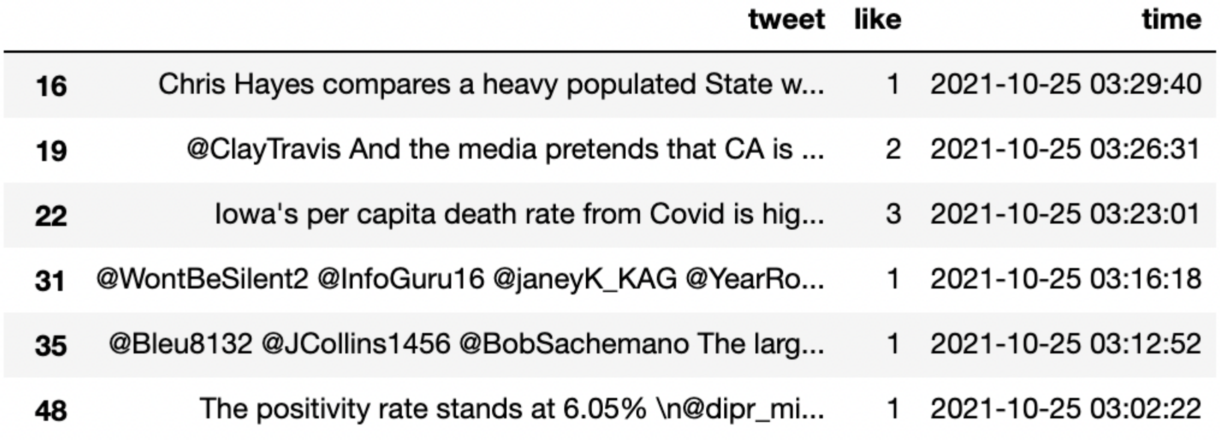
```
#extracting lockdown data from the cursor
for x in lockdown:
    lock_tweet.append(x.full_text)
    lock_like.append(x.favorite_count)
    lock_time.append(x.created_at)
```

Cell Block 8: A data frame was created through the use of pandas library and the skeleton structure made earlier on. Just like the previous step, this same line was repeated throughout the other hashtags as well.



```
#Dataframes
df_mort = pd.DataFrame({'tweet':mort_tweet,'like':mort_like, 'time': mort_time})
df_wfh = pd.DataFrame({'tweet':wfh_tweet, 'like':wfh_like, 'time': wfh_time})
df_job = pd.DataFrame({'tweet':job_tweet, 'like':job_like, 'time': job_time})
df_vac = pd.DataFrame({'tweet':vac_tweet, 'like':vac_like, 'time': vac_time})
df_lck = pd.DataFrame({'tweet':lock_tweet, 'like':lock_like, 'time': lock_time})
```

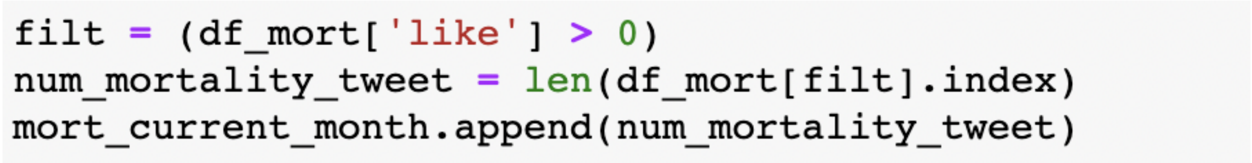
Upon executing the first data frame we get the following output:



	tweet	like	time
16	Chris Hayes compares a heavy populated State w...	1	2021-10-25 03:29:40
19	@ClayTravis And the media pretends that CA is ...	2	2021-10-25 03:26:31
22	Iowa's per capita death rate from Covid is hig...	3	2021-10-25 03:23:01
31	@WontBeSilent2 @InfoGuru16 @janeyK_KAG @YearRo...	1	2021-10-25 03:16:18
35	@Bleu8132 @JCollins1456 @BobSachemano The larg...	1	2021-10-25 03:12:52
48	The positivity rate stands at 6.05% \n@dipr_mi...	1	2021-10-25 03:02:22

The numerical order of the tweets being printed in this data frame for ‘Mortality’ skip a few tweets in between due to the filter that was created. The filter is deployed because a majority of the data that was extracted unfortunately was infested with bots promoting tweets that had nothing of relevance to the topic being investigated.

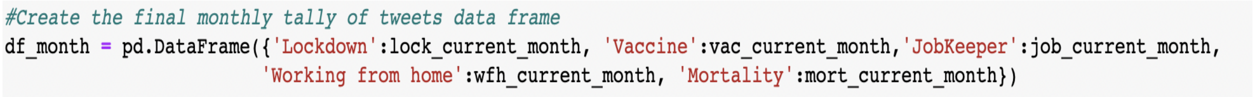
Cell Block 9: The filter variable was created to remove some of the noise in the data being channeled to the data pipeline. In doing so we are able to select tweets that have likes greater than 0 since most ‘botted’ interactivity fails to have any likes higher than 0 making tweets greater than 1 like more authentic. Ultimately increasing the validity of our data set.



```
filt = (df_mort['like'] > 0)
num_mortality_tweet = len(df_mort[filt].index)
mort_current_month.append(num_mortality_tweet)
```

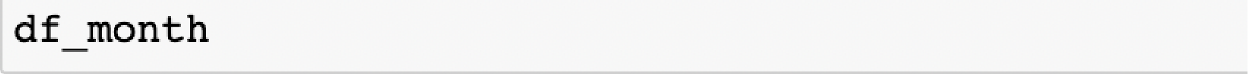
The total amount of monthly tweets is computed through the use of the len() function to determine the amount of filtered tweets in the data frame. The amount of tweets monthly are collected for all other hashtags to create the final data frame.

Cell Block 10: The total monthly tweets are all collated and presented within the same data frame to create the final data set for the monthly performance for each five hashtags that were studied.

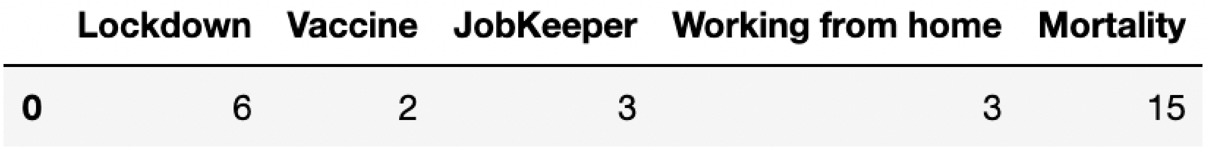


```
#Create the final monthly tally of tweets data frame
df_month = pd.DataFrame({'Lockdown':lock_current_month, 'Vaccine':vac_current_month,'JobKeeper':job_current_month,
                         'Working from home':wfh_current_month, 'Mortality':mort_current_month})
```

In this monthly data frame, 100 tweets where shortlisted for every hashtag when creating the data frame above as the output follows below:

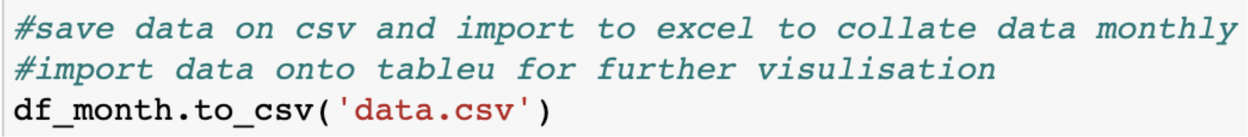


```
df_month
```



	Lockdown	Vaccine	JobKeeper	Working from home	Mortality
0	6	2	3	3	15

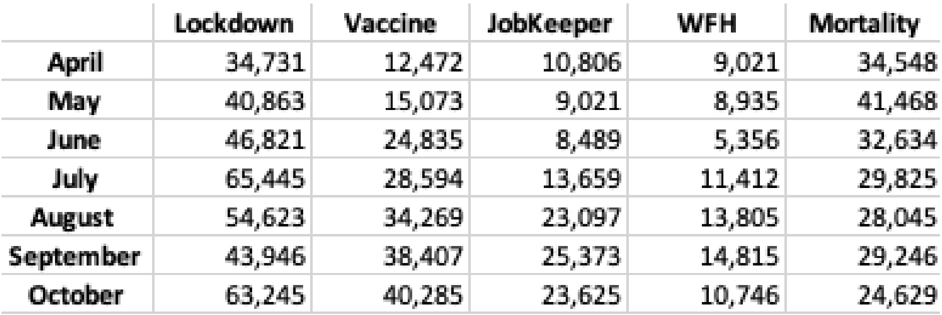
The data was then imported onto a excel file:



```
#save data on csv and import to excel to collate data monthly
#import data onto tableu for further visulisation
df_month.to_csv('data.csv')
```

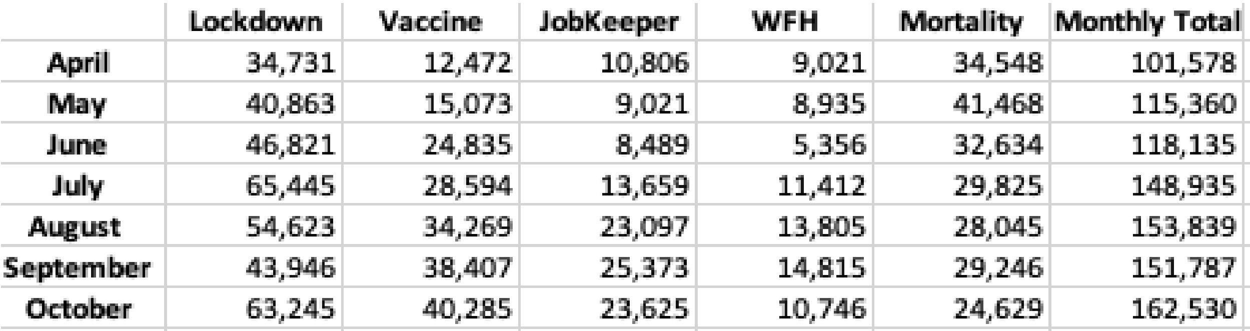
3. Visualisation and Analysis

The data pipeline was autonomously executed once every week to obtain data that was filtered through the code to account for the total tweet activity. Every three weeks the data that was maintained within the data frames were exported onto a csv file that hosts all the data in a table format. This was repeated continuously from April to the end of October before any real patterns could be observed earlier on in the year. The data.csv file was imported into excel primarily to ensure that the format of the data was correct before exporting into tableau. Within 'Excel' the data was re-formatted to account and calculate different visualisation techniques that could be drawn. Initially the data set was hosted in the same manner it was naturally delivered from the data frames utilised. Most of the visualisation that were going to be created was based off this finalised data set:



	Lockdown	Vaccine	JobKeeper	WFH	Mortality
April	34,731	12,472	10,806	9,021	34,548
May	40,863	15,073	9,021	8,935	41,468
June	46,821	24,835	8,489	5,356	32,634
July	65,445	28,594	13,659	11,412	29,825
August	54,623	34,269	23,097	13,805	28,045
September	43,946	38,407	25,373	14,815	29,246
October	63,245	40,285	23,625	10,746	24,629

Within another excel sheet in the same data.csv file the total monthly tweets was calculated by adding the total of all cells in a month to provide a monthly performance of all engagements of tweets mined by the data pipeline:



	Lockdown	Vaccine	JobKeeper	WFH	Mortality	Monthly Total
April	34,731	12,472	10,806	9,021	34,548	101,578
May	40,863	15,073	9,021	8,935	41,468	115,360
June	46,821	24,835	8,489	5,356	32,634	118,135
July	65,445	28,594	13,659	11,412	29,825	148,935
August	54,623	34,269	23,097	13,805	28,045	153,839
September	43,946	38,407	25,373	14,815	29,246	151,787
October	63,245	40,285	23,625	10,746	24,629	162,530

The data was then finally imported into tableau where all the relationships between these hashtags could be observed. Through this, a simple website created through the use of Jekyll being a 'Ruby' based operating language in hosting all combined and individual graphs for each hashtag is hosted. To view the rest of all the visualisations made for this project it can be found here:"https://kosh725.github.io/" . All of the code used within this project within the creation of this site as well as the code present for the data pipeline and the final data.csv generated can be all found within the github repository.

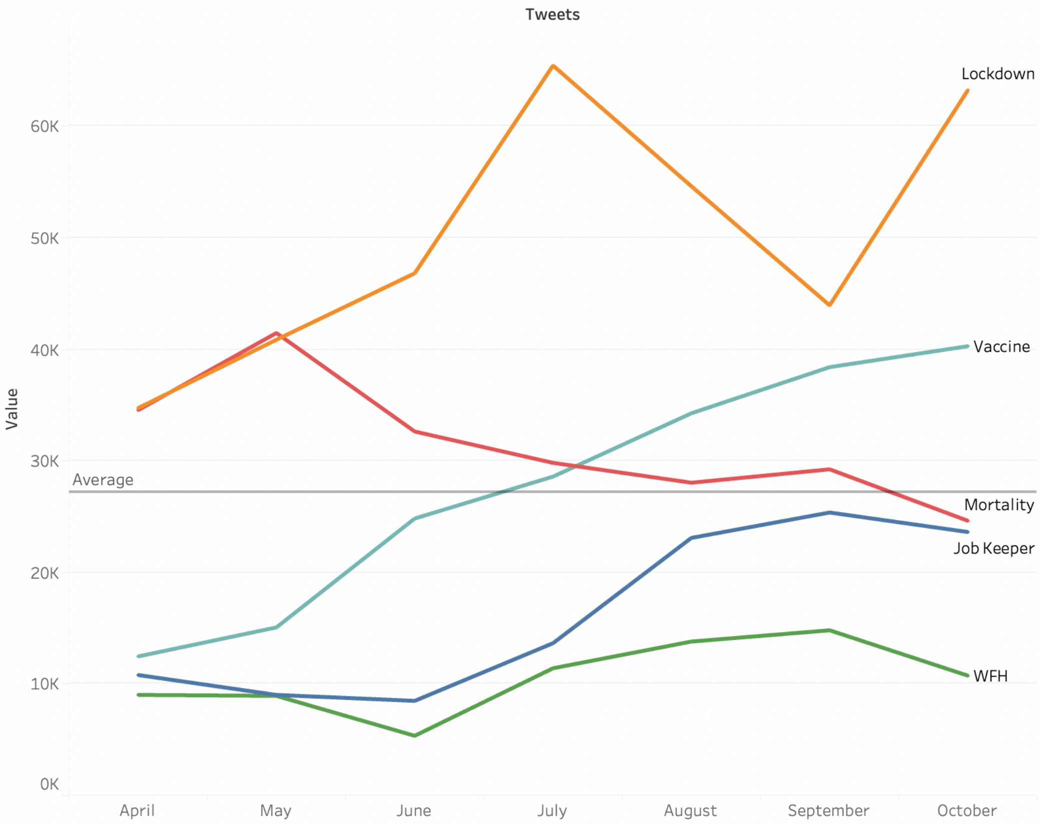
The table below showcases every technological architecture used to help achieved this goal of the capstone project:

Technological Implementation	Use
Jupyter Notebook	The standard data science de facto in creating data pipelines. With plenty of utility and resources being used within this lightweight text editor, it makes for data manipulation to become either with constantly being able to see the output of each cell without the need of constantly executing a terminal to do so. Multiple in built in packages that allow for the import and exportation of dataset to be manipulated within.
Python	Most proficient language used with creating data pipelines. Multiple libraries that are used daily to manipulate data are an extension of python. Pandas which was used within this data pipeline allowed for data to be manipulated by creating succinct data frames. Tweepy API server is purely python based making it mandatory to be used within the data pipeline.
Excel	Used to smooth out any inconsistencies in the data set generated by the data pipeline. Used to also make further data sets that take in the monthly sum of all totaled user tweets.
Tableau	Visualisation program used to create graphs to understand the relationship of the variables within the data set that cannot be identified quantitatively.
Jekyll/Ruby	An old programming language used to host the remainder of visualisations not currently present within the project report. Jekyll synergies relatively well to present static content and data which was needed in publishing the results on of the visualisations publicly.

4. Analysis

The suitable action to measure the performance of each hashtag was to create a line graph to gauge the monthly tweets collated. In this graph below illustrates the monthly total of tweets made about COVID-19 having those relevant hashtags stacked against each other.

Combined: Line Graph



At the start of April the hashtags for “Working from Home” (WFH), “Job Keeper” and “Vaccine” were all relatively close to one another initially. Most of the first world countries at the start of April were out of lockdown which started relatively high at 34,731 tweets which symbolised a lot of mixed feelings about the whole lockdown procedure being gradually lifted. This is reinforced through the monthly total tweets of mortality trailing relatively close to it totalling 34,548 tweets. A catalyst that contributed to the high tweet activity for both these hashtags was the “Delta Variant” strand that swept India in late December 2020 before it started hastily making it through the rest of the world.

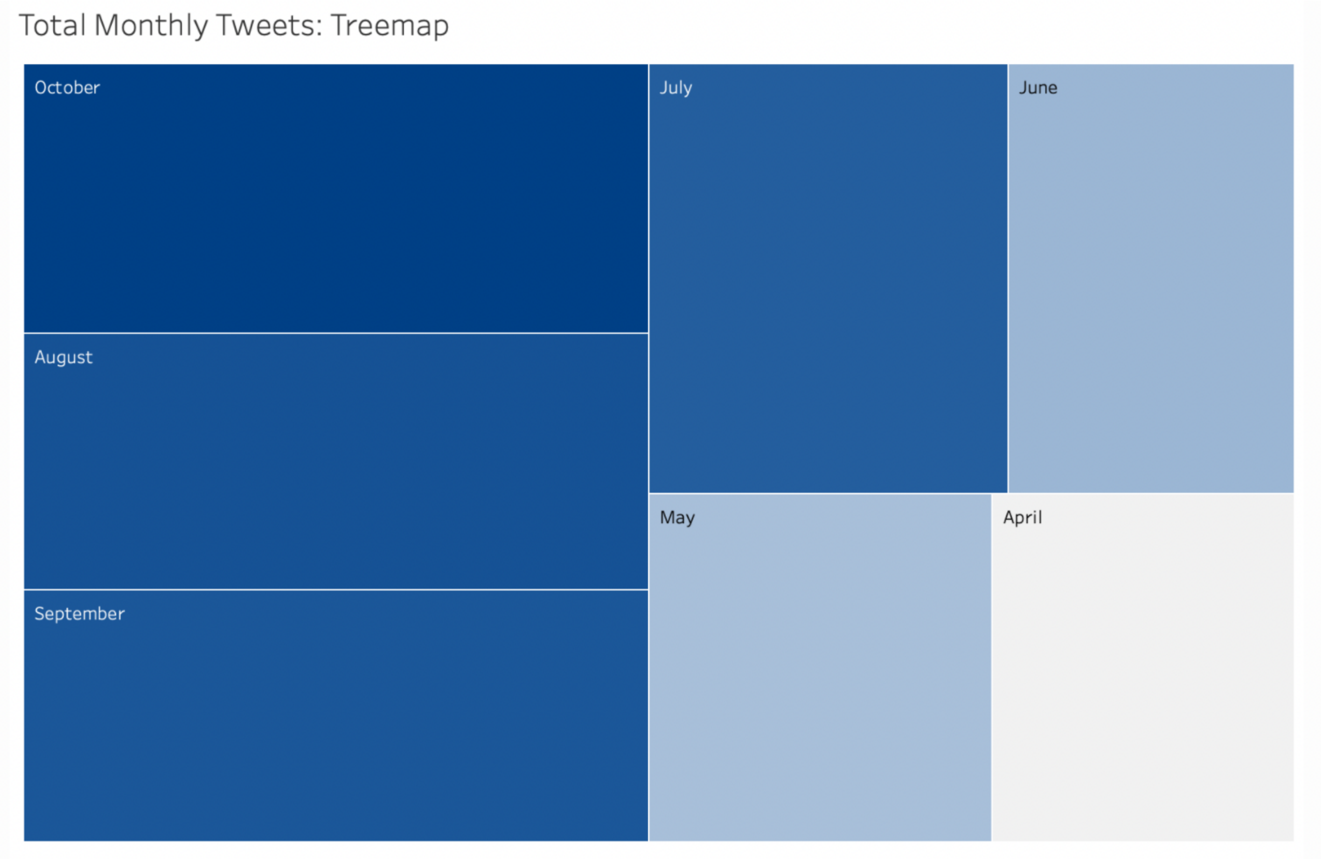
Within May to June lockdown activity had increased by 15% as the delta variant had found its way back to the mainstream populace of first world countries that had lifted a lot of the rules that was enforced against the first strand of COVID-19 in 2020. This in turn ignited the use of the vaccines hashtag being used more by 39% as individuals on the platform were inquiring about the use of vaccines to help combat the prevalent threat. When the vaccine hashtag saw a surge in its use, the mortality hashtag dropped off significantly by 21% with Job Keeper and WFH currently stagnating without any real noticeable changes.

The following months in June to July proved to show the most significant changes of all hashtags as the majority of all countries in the world struggled to contain the Delta strand as effectively. The Australian Health Government had published that a total of 28,408 positive COVID-19 had been recorded in 2020, however during the peak of the delta variant activity within June to July, Australia recorded daily new highs of 1,000+ cases during this month which is on pace to beating last year's total cases that spanned within year in a month. Naturally, Australia alongside other countries including America and Canada imposed a second mandatory lockdown encouraged by the UN as the total number of tweets had increased tremendously by 28.5%. Businesses started to feel the burden of COVID-19 yet again impacting their chances of returning a profit from a previous stagnant year as they started imposing remote work seen through the WFH hashtag that nearly doubled the amount of tweets from the previous month with an increase of 48%. The Australian Government realised that not all business were going to successfully turn over profit so the Job Keeper hashtag saw a slight surge as business on the verge of being liquidated were subsidised by the government to help provide back to the community currently struggling to live through the pandemic as a majority of the population had their work temporarily paused. Those who had lost their jobs were also subsidised through the scheme to ensure that the majority of those who relied on a steady income to get by were also looked after. The vaccine hashtag began increasing so did the decline of mortality hashtag slowly dropping and eventually getting surpassed towards the end of the month. As more users on the platform began voicing their concerns about waiting for their respective governments to receive more doses of vaccines to counteract the delta vaccines.

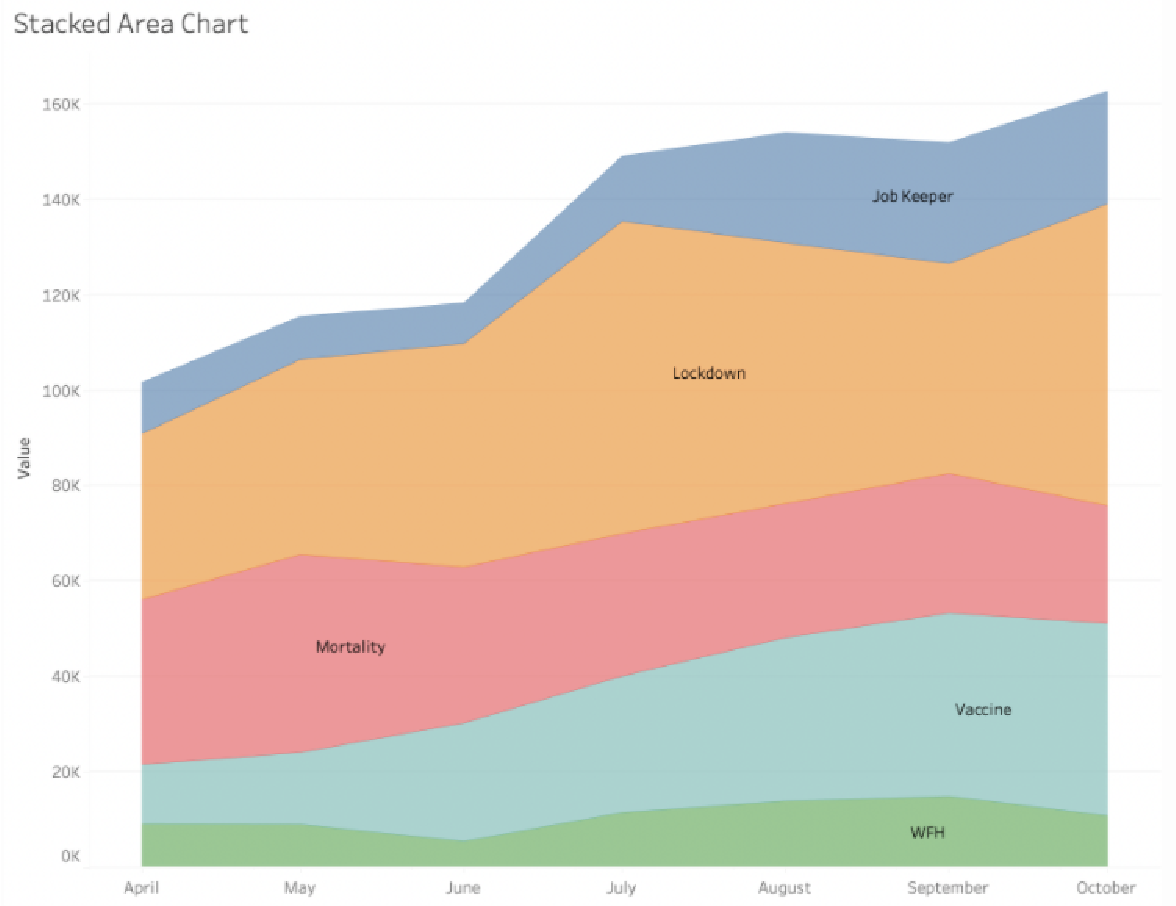
The same trends observed in the previous months were continued to a greater extent however the only outlier to this was the gradual decline of the lockdown tweets. Majority of first world countries allowed citizens to have more relaxed rules placed upon them if they were vaccinated which saw a negative gradient drop of 16.5%. Naturally because of the relaxed rules the activity of vaccines continuously increased without a single drop throughout the previous months as more engagement was created on the topic. The impressions raised within this month for vaccines was enough for the hashtag to surpass the mortality hashtag for the first time in months. It is also within this month that mortality reached its lowest engagement with 20,045 tweets in this month of August as WFH increased slightly by 2393 tweets. Job Keeper saw a dramatic increase of 41% as it was published by the Australian Government that there were a few businesses that were taking advantage of the extra revenue provided to them to stay active during the pandemic when they were more than financially stable. This sparked controversy within Australia as business owners who did not survive the first instance of the pandemic were forcibly shut down to make ends meet whilst the government continued funding businesses who were already financially stable.

During the final recorded month of September to October, the Australian government had announced plans to remove lockdown restrictions during their 'RoadMap' to reopen the country again. It is through this that the lockdown hashtag had a boost of 31% which in turn played a factor in the increase of the vaccine hashtag. Gradually increasing and never steadily dropping over the past months from April, Vaccine impressions on the platform saw a slight rise of 4.7% as a major factor in re-opening the country resulting in the population being vaccinated. Individuals who wanted laxed rules to be imposed on them had to be double vaccinated before the end of October to be able to live any semblance of a normal life post COVID-19. Mortality reached an all time low as the majority of countries around the world have been successfully able to mitigate the death rates in their respective countries. Outlier's to this notion are Ukraine, North America and the United Kingdom.WFH and Job Keeper saw gradual decline yet again as business started to slowly re-open and cater towards individuals who were vaccinated slowly allowing them to generate cash flow within the business. With businesses functioning normally, workers were recalled back into their jobs abiding by the new imposed mandates which eventually tanked the WFH hashtag by 28%.

The treemap map illustrates the ratio within rectangles which month had the most monthly engagement for the total amount of tweets.



April to May had the least number of tweet activity as countries were either slowly focusing on their recovery when dealing with the new delta strand that got introduced into their economy shortly after the first pandemic had ended. June to July saw the slight increase of impressions on the platform as countries attempted to fight off the variant strand. October had the most number of monthly tweets as it was the start of the country slowly recovering and re-opening from the damages done through the delta strand. August and September were the trends of the countries globally recovering starting manifesting that started to flourish within these months.



This stacked area chart further showcases the amount of monthly tweets totalled by each hashtag. Lockdown was responsible for having the most amount of engagements on the platform each month which can be seen by its constant growth apart from the slow hit it endured within July to September. Mortality and Vaccines almost have an identical mirror pattern that further illustrates the relationship between one another. As the usage of mortality increased vaccines slowly followed. However as lockdown gradually grew bigger mortality engagements drastically increased allowing vaccines to flourish throughout the end of the graph. More noticeable trends also include WFH and JobKeeper being heavily influenced by the lockdown trend. The more relevant lockdown became relevant which can be seen by its drastic spike within June, WFH and Jobkeeper grew significantly and are still maintaining its significance throughout the forthcoming years.

5. Discussion

There were a few limitations that were encountered when conducting this project throughout the years. “Tweepy” the API server that was used to collect and mine the majority of the data presented the most difficulty in helping accurately capture the nature of the hashtags studied. Before the commencement of creating a twitter developer account to interact with tweepy, there were a few other hashtags that were initially proposed such as “Stimulus Check”, “Hospital/ICU”, “Jobs” and “Coronavirus”. However, as relevant as these hashtags were throughout 2020 and early 2021, Tweepy does not grant access to tweets that are older than a week making it impossible to study these particular trends. A large factor of understanding the relationship of social interactions on online platforms was taken away as the height of the pandemic could not be studied and in turn could not be visualised to a certain extent. When initially requesting the data to test the volatility of these hashtags, they received little to no attention even when the ramifications of these trends are still felt today. Furthermore, Tweepy does not allow requests to be made based on geolocation of a particular country. When initialising the cursor to draw tweets, the only parameter that can be filled to take in a particular location is coordinate based meaning that only a specific area of a city can be studied.

When studying the meta-data presented within the directory of each tweet, the majority of users opted out of providing their current location on their tweets making it redundant to use any visualisation techniques based on geolocation to draw heat maps. Geographic/Heat maps are very powerful visualisation techniques that allow you to observe the behaviour of user’s worldwide and it is an opportunity missed to capture this behaviour during the pandemic. In understanding all these current constraints being placed on the nature of this project it resulted in allowing the second wave of the pandemic to be observed throughout the world. All hashtags that were used in this project were the most popular trends to be observed as other lesser trends could take away from them. These hashtags weren’t chosen because Tweepy's final limitation that took away from the project was being restricted to 500,000 tweets monthly. Initially it sounded plentiful, however when filtering through the data set there were thousands of tweets that were discarded due to them simply being ‘botted’ activity. Majority of the filters created followed the patterns made by these bots such as selecting tweets with more than 1 like and no eccentric characters. Unfortunately there were some tweets that were created by authenticated users which did follow these circumstances and were discarded along the way but ultimately increased the validity of the data set. This was the main highlight as once a particular hashtag grew popular so too did the amount of automated bots using these hashtags to convey opinions on something entirely different became. The fewer hashtags observed allowed for more bandwidth to be used from the half million tweet limit restriction. These factors contributed in studying more current trends as this project was able to further explore the dynamic of the delta variant strand throughout this year.

Initialising the pipeline came with a few challenges as it became quickly convoluted throughout the jupyter notebook instance. When the github repository was created, the access tokens were mistakenly published on it for a couple of weeks allowing it to be viewed by others. This resulted in some strange patterns as someone else had started using my access token to enter my account and change some settings that had restricted me from accessing tweepy earlier in the year. Upon realising the gravity of the situation, the access tokens generated were immediately destroyed and all the mandatory permissions requiring my password use were also indefinitely changed. Overlooking the security aspect of the pipeline had cost thousands of tweets to be currently lost that were being stored in the current session as they expired due to the fact that the tokens that were used to access them no longer existed. The tokens that were used to create an authentication access to the Tweepy servers were hosted on a local jupyter notebook file that was run simultaneously alongside the data pipeline. The keys were imported onto the main pipeline and were executed before any mining had begun. Utilising this method prevented any public and private access tokens being exposed.

A particular cell was responsible for tracking all 5 hashtags at the time before realising how inefficient this process was simply because executing this cell often led to multiple run time errors. Majority of the cursor initialisation was processed together resulting in almost half a million tweets being extracted within one cell making mining tweets a lot longer than it had initially planned to take. The cell was optimised by making it modular in separating different hashtags to newly created cells allowing for further control on the amount of tweets that needed to be extracted within a particular week. Gaining a higher retention of control over the pipeline itself generates more accurate results within the dataset.This almost indefinitely solved the run time errors being delivered and it increased the speed of the pipeline being run weekly and autonomously without little effort.

Tableau was the visualisation program that helped elevate the understanding of the data set by creating multiple different graphs. Each visualisation depicts the data set in a different light and it is up to the interpretation of the reader to gain a more insight within it. However, the proficiency needed to create far more complex visualisations is solely dependent on the amount of variables within the data set. Within this data set, five variables were measured against one another in months resulting in somewhat rather complex visualisations. Other visualisations could not be made simply because the data required for more visualisation require more dimensions to explore it further. Creating a multi-dimensional data set was impossible as the data that was being tested before the data pipeline was created showcased a lot of inconsistencies that made it near redundant to continue chasing. It is through this that the quality of the visualisations had been impacted to a certain extent.

6. Conclusion

Throughout the year, each hashtag that was monitored over this seven month stint played a factor in understanding the hidden relationships between one another. Each hashtag that was studied measures the populace response to certain events that occurred and within this window the delta strand saw the rise and fall of particular trends. This was carefully illustrated through the use of various visualisation techniques that showcased this notion. As the delta variant became more of a global threat the mortality hashtag saw some slight growth that rivaled the same amount of growth lockdown had as user's began exercising this hashtag to express their opinions. This particular instance also gave rise to the slight attraction that the vaccine hashtag was being received as the current trends show that both had increased relatively by the same amount. However, as more users began picking up the vaccine hashtag as a response to the growing threat of the delta variant more users became reluctant to use the mortality hashtag as it was slowly phased out. The gradual regression of mortality allowed the rise for the vaccine hashtag to thrive further which surpassed the mortality hashtag a couple of days into June. The mortality rate of COVID-19 had also started to be slowly controlled through more vaccines being received.

Almost all countries that had access to the technology and resources to be able outsource vaccines into the populace allowed them to effectively fight off the delta variant to a certain extent as it was initially met with resistance. Within the first few days of June, we observed the highest peak of the lockdown hashtag recorded as it also was used as a response to help mitigate the virus which in turn allowed the "JobKeeper" and "Working From Home" hashtags to inevitably be used more frequently as a consequence. When lockdown mandates became more attractive so did the need to express support from the government to survive in an uncertain pandemic lifestyle.

Further research can be implored within this project as visualisation was only the first step in truly grasping the contents of the dataset studied throughout this year. A sentiment analysis can be conducted and modelled through the data set mined to understand online social behaviour driven by real life instances throughout the tweepy API. Utilsing a sentiment analysis requires a model to be created based on a training data set. It is crucial to use a reliable training data set that allows the model to be able to accurately predict a variety of emotions. Reputable data sets that can be used can be sourced throughout 'Kaggle' as majority of datasets have been thoroughly vetted by renowned scientists within the fields. Recurring keywords that reflect human behaviour used throughout the tweets can be tabulated and calculated to reference a human emotion to create a test data set that can accurately predict the sentiment of users online during the pandemic. Deploying this model will be able to be referenced throughout this current pandemic and also to continuously learn throughout other real time events happening throughout the world. Visualisations can be created to model the wide variety of emotions of users online to understand the overall sentiment of the populace's interactions with said event occurring around.

By carefully studying previous scholarly articles that helped solidify the importance of basic data visualisation principles and implementing the "Data Science Methodology" were crucial factors in providing an in-depth analysis on the results obtained. The project aims have been answered through the use of visualisation programs such as Excel for smoothing out the data set and Tableau providing multiple interpretations of the five hashtags that were set out to be studied thoroughly throughout April to October.

7. Bibliography

- Katella, K. (2021, July 30). *5 Things To Know About the Delta Variant*. Yale Medicine. https://www.yalemedicine.org/news/5-things-to-know-delta-variant-covid
- Tsao, S.-F., Chen, H., Tisseverasinghe, T., Yang, Y., Li, L., & Butt, Z. A. (2021). What social media told us in the time of COVID-19: a scoping review. *The Lancet Digital Health*, *0*(0). https://doi.org/10.1016/S2589-7500(20)30315-0
- Telea, A. C. (2014). Data Visualization: Principles and Practice, Second Edition. In *Google Books*. CRC Press. https://books.google.com.au/books?hl=en&lr=&id=AGjOBQAAQBAJ&oi=fnd&pg=PP1&dq=data+visualization&ots=NIxuvo1XBI&sig=tKX1U9fCYx8wWSgP_77InU2J-zY&redir_esc=y#v=onepage&q=data%20visualization&f=false