The Relationship(s) between Australian Sentiments, Unemployment Levels, and the Stock Market.

By Tianzheng Tong (1025169), David Capes (1082166), Ryan Ashe (1608327), and Kai Ikeda (1531590)

Aim(s):

The general question we aim to answer is; what is the relationship between Australians’ moods, unemployment levels, and the stock market? We chose these three disparate, yet related metrics with the hope that we could uncover some interesting patterns and relationships.

Our question can be more precisely reframed as three distinct questions:

1. What is (if any) the statistical relationship between the evaluated sentiments of Australian tweets and the daily ASX 200 from mid-2021 to mid-2022?
2. What is (if any) the statistical relationship between the evaluated sentiments of Australian tweets and the unemployment levels across various states during mid-2021?
3. What is (if any) the statistical relationship between the overall unemployment level across Australia and the ASX 200 from 2012 to 2021?

Our expectation/hypothesis is that only large disturbances to employment levels (such as state-based recessions) would be able to have a noticeable effect on either ASX or Twitter data. Similar to the ASX 200 data, we only believe it will have much impact during large market fluctuations. However, it is plausible that such large disturbances could have occurred in the given periods. Ultimately, however, the aim/goal of this project is more than simply the question of the investigation itself, but to build a scalable cloud-based system that can process and analyse these datasets.

Technologies (Background):

In recent years, cloud-based parallel computation has become the industry standard for big data analysis. For this project, we make use of a plethora of complementary technologies including but not limited to Kurbernetes, Fission, and Elastic Search all on the MCR. We will further discuss the details; including functionality, advantages, and disadvantages of these various technologies.

Kubernetes

Kubernetes is an industry-standard open-source platform that enables the automation of deployment, scaling, and management of containerized applications. Kubernetes as a whole is an incredibly complex system, however, for this task, our usage of it was far more limited.

One of the strongest advantages of Kubernetes is its scalability. Automatically scaling enabled the application to remain at peak performance and can quickly respond to increased traffic. However, in this, there wouldn't necessarily be an increase in traffic to any of the nodes requiring the deployment of additional nodes on the cluster. As a result, this project utilised Kubernetes clusters to distribute the workload evenly allowing for high availability and protection against faults. In the case of a node failure, Kubernetes would automatically restart, replicate, or reschedule the node. Reducing downtime and maintaining continuous operation maximises the efficiency of data processing.

Another advantage of Kubernetes that we employed in our assignment is the use of declarative configuration files to manage the infrastructure. While it is possible to deploy the Kubernetes straight into the command line, the usage of that would lead to work being incredibly messy and hard to track. Utilising configuration files ensured consistency and simplified the management of Kubernetes by setting the desired state, and controlling it from the configuration file. The configuration file also enabled Kubernetes to continuously reconcile the actual state of the nodes with the desired state in the configuration file.

Kubernetes also has the significant advantage of portability which was employed for this project. Because of the underlying structure of Kubernetes, it can be deployed across various environments, like in the case of this project by being deployed on the Melbourne Research Cloud. Additionally, due to Kubernetes being the industry standard for container orchestration, there is a vast array of extensions and tools that can enhance its functionality. Monitoring, logging, security, and networking tools can all be integrated to meet the unique demands of unique situations where Kubernetes is employed. The Kubernetes community provides a wealth of online resources including documentation, tutorials, forums, and videos. This allows troubleshooting issues to be quickly resolved through leveraging the collective knowledge of other kubernetes users. These aspects made Kubernetes an incredibly powerful tool deployed in this project.

Fission:

Fission is a serverless framework that was designed to run on Kubernetes. We employed it in this project to deploy and manage functions and not worry about the underlying function. It has many advantages and disadvantages which we encountered with the development of the project. One of the key advantages of Fission is that it is an open-source framework allowing easy access and the ability to modify the code to both understand it better, and to also ensure that it meets our needs for the project. Additionally, as it is a powerful tool that can efficiently manage the resources it allows us to focus on writing functions and developing code instead of focussing on the infrastructure.

One of the significant benefits of using fission in this project compared to other serverless frameworks like Knative is its simplicity. Fission allowed us to streamline the deployment by allowing us to avoid the need to build Docker images. Avoiding the use of Docker allowed us to deploy functions quicker. It is also enhanced by its support for a vast range of programming languages, including Node.js, Go, C#, and what we used for this project, Python. This flexibility to use a preferred programming language for the project makes it incredibly easy to use. Fission also facilitates real-time data processing through message queues, timers and HTTP requests allowing tasks to be handled in synchronisation.

We did encounter issues with diagnosing and debugging the serverless functions due to the complexity of their distribution. Fission did provide us with monitoring tools allowing us to diagnose the issues, however the ability to resolve those issues was still often challenging. Managing the dependencies of the serverless functions was rather difficult for us due to this being our first project in managing cloud based software. It also required extensive research to ensure that we employed and configured all the tools and libraries needed.

The primary issue with Fission is that while it is a simpler serverless framework than Knative it still requires a learning curve in relation to understanding the serverless concepts and integrating Fission with Kubernetes. However, the user-friendly design greatly assisted in overcoming the learning curve.

Fission provided us with a powerful and flexible solution for running serverless functions on Kubernetes, thus making it an incredibly powerful tool for this project. Its ease of use, supporting multiple coding languages made it the optimal tool choice for this project.

Elastic Search:

Elasticsearch was another source of technology we used for this project. Elastic search is an open-source search and analytics engine which has robust capabilities in handling large volumes of data. For our project volume primarily related to its size as we were not using data that involved varying data. ElasticSearch was also used because of its ability to dynamically search all the data for matches of keywords within the tweets contents.

The ElasticSearch architecture is designed for distributed environments enabling efficient data management and querying across multiple nodes. This employed the node distribution of kubernetes and fission allowed for incredibly efficient data analysis. Each ElasticSearch cluster contains nodes, and those nodes are all running instances of elastic search. The JSON file used was split into shards, enhancing the performance to assist in speeding up the analysis. The shards that were created from the data also had replicas, to ensure that there was protection against faults and data redundancy.

In particular ElasticSearch was crucial for our project due to how effective it is in group searching. The nature of our question involved capturing the sentiment from different states, and the use of elastic allowed the clustering of tweets efficient. This made ElasticSearch pivotal as it was able to efficiently search through the large volume of tweets, capturing the important data we needed for our analysis. Additionally its ability to use a group by sort on the unemployment database was incredibly helpful. The unemployment database did not contain states, and instead only regional codes for certain local government areas. Utilising the group by sorting all the data into the state categories allowed our front end analysis to be conducted efficiently. While we did not use it for geospatial data, its ability to also look at geospatial data makes it incredibly versatile for different applications, and would be critical for data analysis in even larger and more complex datasets.

The weaknesses of ElasticSearch we were able to avoid due to the nature of our queries. ElasticSearch falters in more complex relational data processing, and is not as robust as traditional relational databases in that way. As we were only looking for specific query questions we were able to adapt the ElasticSearch to fit our needs in relation to the datasets, avoiding its weaknesses. Additionally, because the dataset that we were analysing was not fluctuating in size we were able to explicitly map the data, which strengthened the usefulness of ElasticSearch. While ElasticSearch has the ability to dynamically map and adapt to documents, its performance and efficiency is enhanced when explicitly mapped.

One of the most advantageous components of the ElasticSearch ecosystem used would be the kiana interface. It allowed us to interface and visualise the data that was being managed. Additionally, it assisted in us looking at transformations that we could make on the data. The additional tools that we could integrate into ElasticSearch made the use of the technology easy, providing an effective solution for data ingestion, processing, and visualisation.

Melbourne Research Cloud (MRC) and OpenStack:

The Melbourne Research Cloud (MRC) is a distributed cloud computing system operated by the University of Melbourne that we made use of during this project. The MRC runs on the OpenStack framework, though it does not have access to the full range of services available in OpenStack. This is because the MRC only contains a subset of what OpenStack provides, and of that subset, we were only allowed to use a further subset of the MRC's services.

The MRC has many advantages and disadvantages, both when compared to traditional computation methods, and other cloud-based systems. The primary reason we made use of the MRC here is due to its availability. For our subject (COMP90024), we were allocated 11 cores and 700 GB of storage on the MRC. Comparatively, it is often prohibitively expensive to use commercial cloud computing services such as Google Cloud Platform (GCP), Microsoft Azure, and Amazon Web Services (AWS). However, for wealthy individuals/entities, the process of obtaining nodes on these commercial systems is far easier than MRC, and can facilitate mid-project upscaling far better. These commercial systems also tend to be far easier to use than OpenStack-based systems like the MRC. Their variety of management services and interfacing drastically reduces the system set-up time and complexity, allowing more time for focusing and tuning other aspects of a project. The trade-off for this useability is that OpenStack systems tend to be a lot more flexible in what they can accomplish, and the MRC is an example of this. However, as noted earlier, MRC doesn’t make use of all that is available in OpenStack.

For instance, until recently Kubernetes was not available on the MRC, and was only added when demand was great enough, which could have been limiting for researchers wishing to use it prior. This rigidity is a drawback that applies specifically to the MRC but not generally to OpenStack systems. There is also infrastructure present in commercial systems that OpenStack does not have. Examples of this are Managed Hosting Platforms (Elastic Beanstalk), Content Distribution Networks (CloudFront), as well as some of the services that are present but less developed, for example, Heat vs. AWS Cloudformation in the aspect of autoscaling.

Cloud computing infrastructure evolves at a very fast pace, with many, even recent technologies being left behind or rendered obsolete. No one, us included can predict the future, however, a general trend in the field of emerging technologies is that systems that are more user complicated; even to the benefit of flexibility, tend to be left behind. An example of this is how many people choose to program in languages like Python now instead of Fortran or Assembly, despite the latter giving users more control over the machine. Other systems like AWS offer greater ease of use and are already larger and thus receive more attention, add-ons, and funding than OpenStack systems. While the future is absolutely not certain in the world of Cloud Computing, these factors are still very important to consider for deciding what software to conduct projects on and dedicate time to. However on the upside, for a changing field, OpenStack's flexible nature also positions it very well.

Datasets:

Twitter Dataset:

The first dataset utilised in this project is the 100-gigabyte Twitter dataset in the form of JSON that was provided during Assignment 1. This dataset contains a vast amount of data gathered from Twitter, offering invaluable insights into how sentiments were being expressed on social media at a given time. For our analysis we will be focussing on a subset of the data relating to the overall sentiment of tweets for all of Australia, comparing it to the ASX data, and also comparing the change in sentiment over a 1 quarter period with the unemployment rate change in the same quarter. The fields retained will include the numerical sentiment score of each tweet, the date and time of tweeting, and the location will not be stored, but instead used by elastic search to thin out the dataset.

While this dataset does encompass a large array of interactions, it only encapsulates a niche subsect of the general populace. The Twitter user base is not a perfect representation of the general population as not everyone uses twitter, and some groups will use twitter more than others. For it's interaction with the ASX dataset there might be less individuals who express their sentiment in regards to the ASX on twitter. Additionally, the frequency of their tweets might be outnumbered by other individuals who tweet far more frequently, skewing the data. The sentiment scores generated also may include errors, as they are derived from language models and could misinterpret the information from the tweet, assigning it an incorrect sentiment. Another area for concern is that sentiment might fluctuate due to completely unrelated factors to what we are looking at, and could induce false positives with our information.

Unemployment Dataset:

Another dataset we utilised came from the Spatial Urban Data Observatory (SUDO). This data displays numerical estimates for the unemployment level across Local Government Areas (LGA) in Australia from December 2010 up to September 2021. The reporting period for this data is quarterly (four times per year), which can provide semi-granular insight into the economic trends of LGAs.

Important considerations regarding this dataset are that they are estimates, and it's difficult to truly know unemployment levels, as unemployed people are harder to track. Though the estimates may be correct, potential systematic biases from each LGA could compound these errors greatly across the population. Some LGAs (particularly across the Northern Territory and Tasmania) have partially or fully incomplete results, and some areas may have much better and more rigorous methods of estimating unemployment. The other key detail dataset is that the frequency is quarterly, which is typically good, especially for fields like employment that typically don't greatly fluctuate day-to-day. Nonetheless, it limits our resolution towards shorter-term trends. The final consideration of this dataset is the absence of population metrics since unemployment is being presented as a raw integer. This is not necessarily an issue, population data could be employed separately and unemployment levels can still be compared and changes observed. But it is a bias we had to consider for the purpose of analysis.

This dataset is being used to see if unemployment has an effect on twitter sentiment and whether the ASX value increases or decreases based upon the rate of unemployment.

ASX Data:

The third dataset we employed for our analysis, was an index that tracks the Australian Stock Market (ASX 200). This dataset reports opening prices, closing prices, minimums, maximums, and percentage changes of the stock prices for each day from 03/01/2012 to 23/05/2024.

A limitation of this dataset, and our analysis of it involves the fact that what influences the shift in stock market performance isn't limited to only the two variables we are looking at here with our analysis. Global economic events, political developments and other economic situations can impact stock prices, completely unrelated to our datasets. This can lead to potential false positives, where we will draw a conclusion from our analysis, even though the correlation did not in fact occur, and it was a situational occurrence.

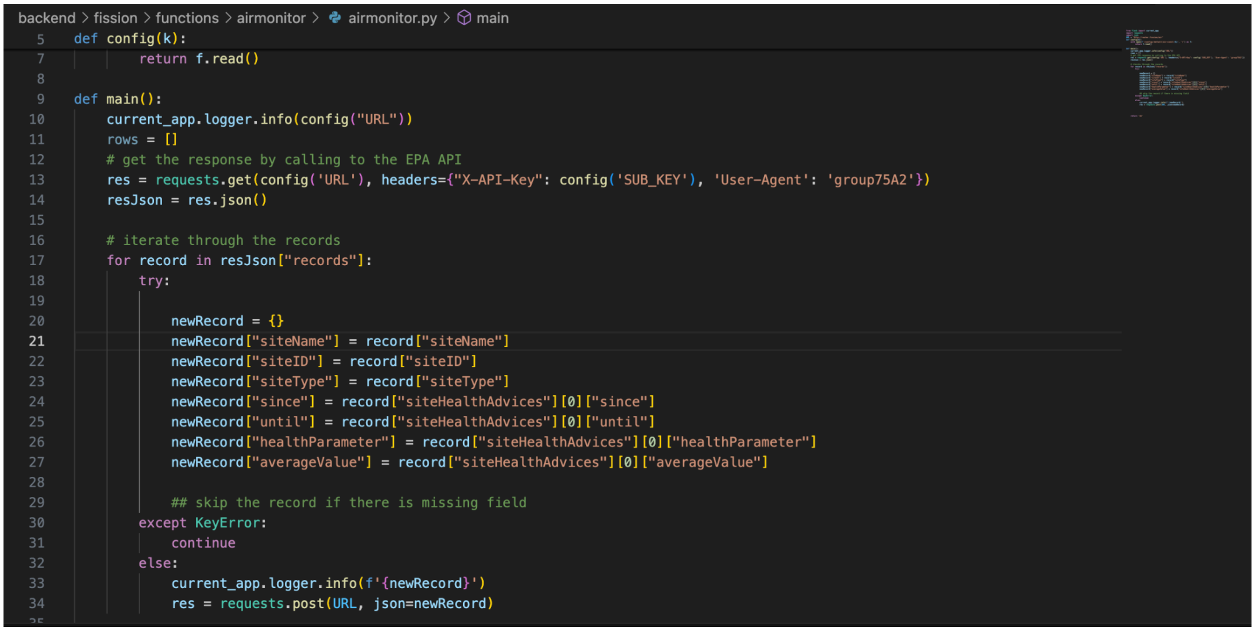
This dataset is being used to see if the ASX value increases or decreases based upon the twitter sentiment and unemployment dataset. Does twitter sentiment affect the daily changes of the ASX dataset, and does quarterly changes in unemployment have a noticeable effect on the ASX changes

System Functionality:

Air Condition Data from EPA:

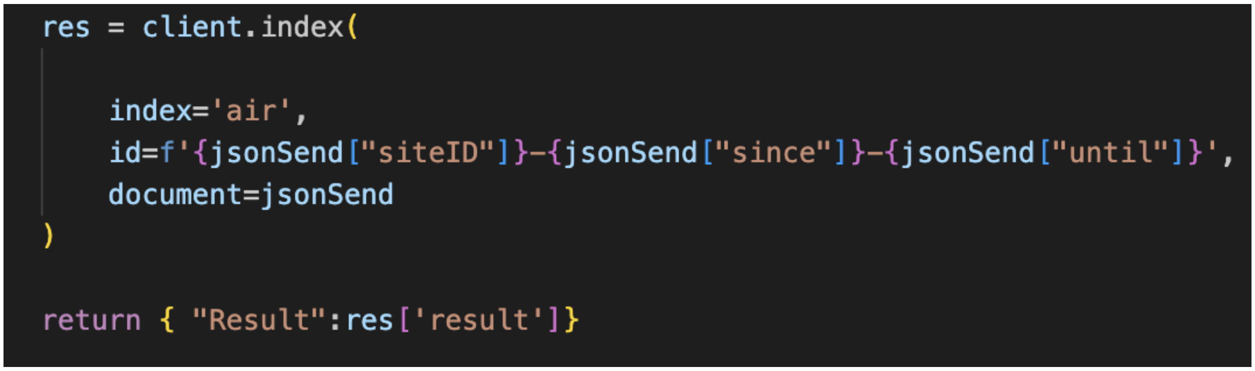
To show the ability of our system to harvest data by calling a remote API, although the collected data from EPA is not analysed in figures and graphs, our system is collecting data hourly with a timer. The dataset contains siteName, siteID, siteType, since, until, health parameter(such as PM2.5 or others) and value for that parameter.

With the help of a timer from fission, a function is created to call the API provided by EPA every hour. After receiving the response in json, the function will parse the json and call another fission function to index each document which is recognized. The url to EPA and api key is stored in a config map. Here is snapshot of the code:


backend > fission > functions > airmonitor > airmonitor.py > main

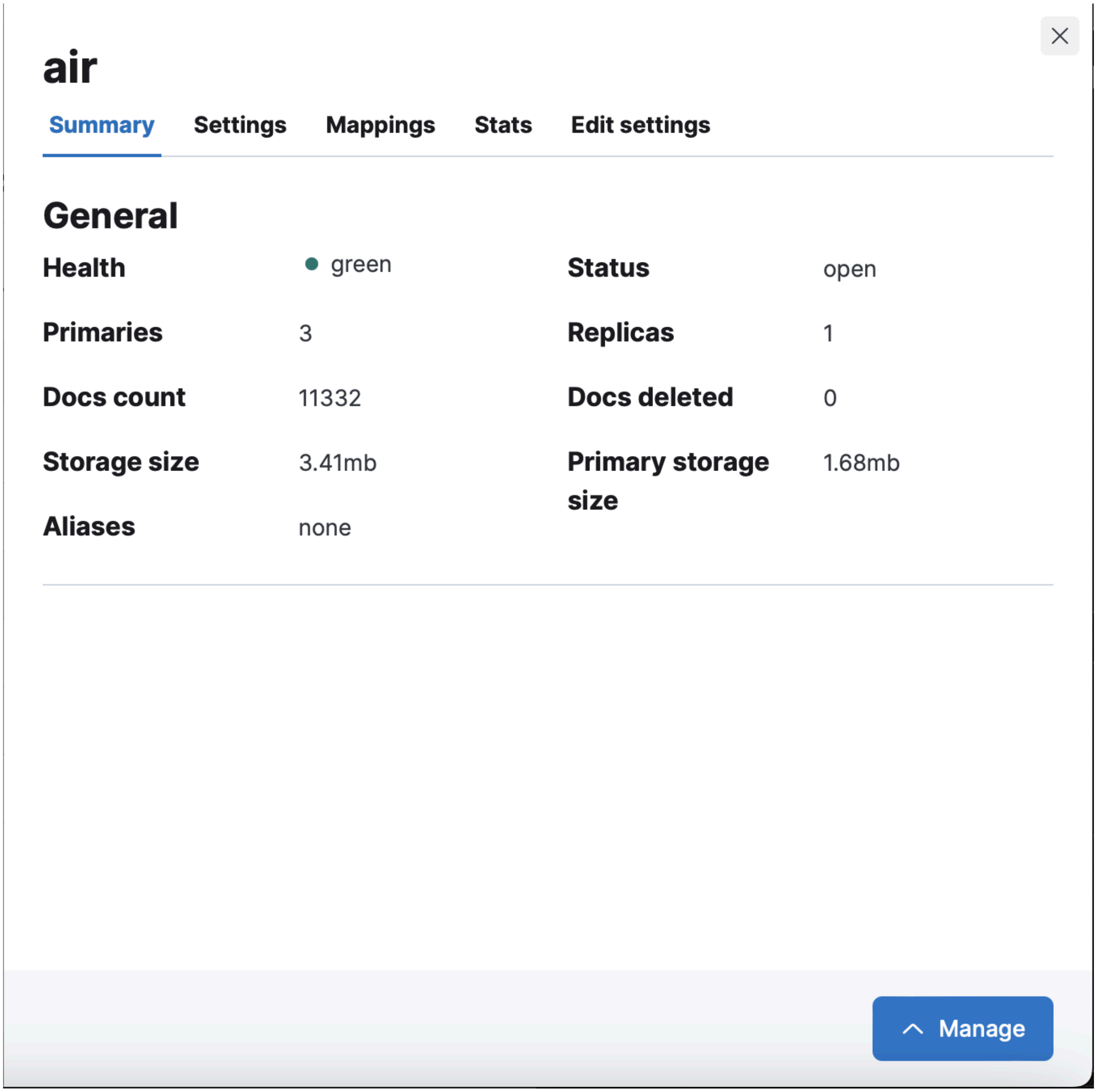
```python
def config(k):
        return f.read()

def main():
    current_app.logger.info(config("URL"))
    rows = []
    # get the response by calling to the EPA API
    res = requests.get(config('URL'), headers={"X-API-Key": config('SUB_KEY'), 'User-Agent': 'group75A2'})
    resJson = res.json()

    # iterate through the records
    for record in resJson["records"]:
        try:

            newRecord = {}
            newRecord["siteName"] = record["siteName"]
            newRecord["siteID"] = record["siteID"]
            newRecord["siteType"] = record["siteType"]
            newRecord["since"] = record["siteHealthAdvices"][0]["since"]
            newRecord["until"] = record["siteHealthAdvices"][0]["until"]
            newRecord["healthParameter"] = record["siteHealthAdvices"][0]["healthParameter"]
            newRecord["averageValue"] = record["siteHealthAdvices"][0]["averageValue"]

            ## skip the record if there is missing field
        except KeyError:
            continue
        else:
            current_app.logger.info(f'{newRecord}')
            res = requests.post(URL, json=newRecord)
```

And the mapping is already created, the indexing process for the other function is easy. It is done by setting the document as the received json and id which combines the siteID and time range of when the data collected.



```python
res = client.index(

    index='air',
    id=f'{jsonSend["siteID"]}-{jsonSend["since"]}-{jsonSend["until"]}',
    document=jsonSend
)

return { "Result":res['result']}
```

After collecting data for a week, it has 11332 documents and is increasing with a speed of 1000 documents every day.



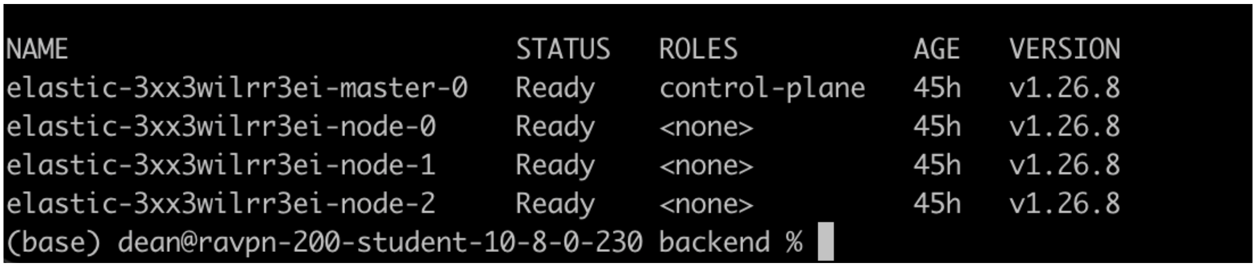
air

General

Health	● green	Status	open
Primaries	3	Replicas	1
Docs count	11332	Docs deleted	0
Storage size	3.41mb	Primary storage size	1.68mb
Aliases	none		

Main System:

The aforementioned categories were employed for our project. In our system architecture, Kubernetes played a pivotal role in orchestrating and deploying containerised applications across the Melbourne Research Cloud (MRC). Kubernetes ensured that the data processing pipeline was scalable, as it was filtering through a Twitter dataset. Kubernetes pushed the workload evenly across the cluster nodes performing elastic search. We employed the usage of 1 master node and 3 worker nodes.

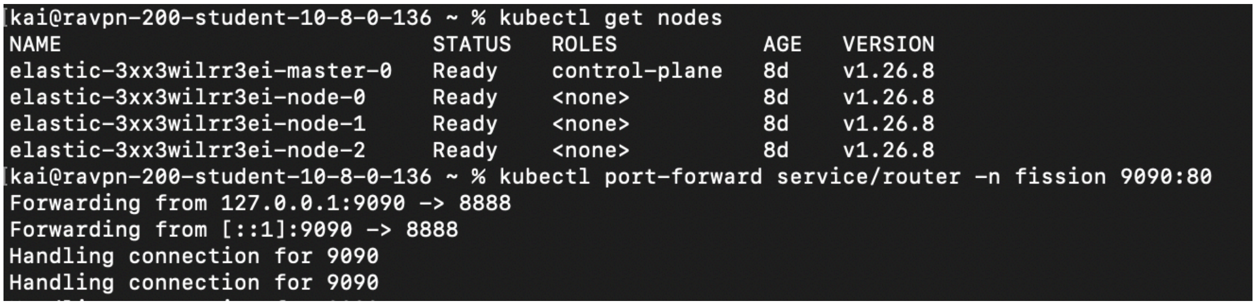


```
NAME                            STATUS   ROLES           AGE   VERSION
elastic-3xx3wilrr3ei-master-0   Ready    control-plane   45h   v1.26.8
elastic-3xx3wilrr3ei-node-0     Ready    <none>          45h   v1.26.8
elastic-3xx3wilrr3ei-node-1     Ready    <none>          45h   v1.26.8
elastic-3xx3wilrr3ei-node-2     Ready    <none>          45h   v1.26.8
(base) dean@ravpn-200-student-10-8-0-230 backend %
```

As can be seen in the above screenshot, this is a display of the Kubernetes clustering of ElasticSearch nodes through Fission. This screenshot was taken while the nodes were inactive and not processing any data through elastic search.

ReSTful design:

How the server architecture is designed in line with restful design
All datasets used in the project are published through the server at 127.0.0.1:9090. These include detailed data necessary for analysis, which is filtered through ElasticSearch according to specific conditions.



```
[kai@ravpn-200-student-10-8-0-136 ~ % kubectl get nodes
 NAME                            STATUS   ROLES           AGE   VERSION
 elastic-3xx3wilrr3ei-master-0   Ready    control-plane   8d    v1.26.8
 elastic-3xx3wilrr3ei-node-0     Ready    <none>          8d    v1.26.8
 elastic-3xx3wilrr3ei-node-1     Ready    <none>          8d    v1.26.8
 elastic-3xx3wilrr3ei-node-2     Ready    <none>          8d    v1.26.8
[kai@ravpn-200-student-10-8-0-136 ~ % kubectl port-forward service/router -n fission 9090:80
 Forwarding from 127.0.0.1:9090 -> 8888
 Forwarding from [::1]:9090 -> 8888
 Handling connection for 9090
 Handling connection for 9090
```

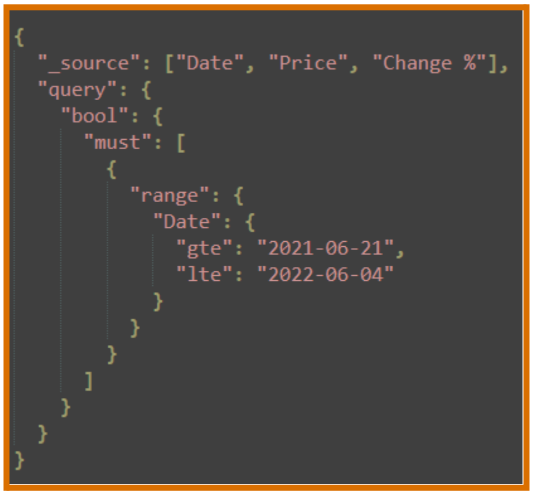
Resource Identification:
Each resource is identified by a URI. In our project, we primarily identify three main data resources (ASX, unemployment, Twitter dataset). Then, adjust the detailed information necessary for analysis using ElasticSearch, uniquely identifying data resources according to conditions with URIs such as 'http://127.0.0.1:9090/asx/quarterly' or 'http://127.0.0.1:9090/tweets/state'.

Uniform Interface:
All resources are retrieved using the GET method, one of the HTTP methods, which is suitable for obtaining information from the server without altering the state of the resource. For example, by executing 'requests.get('http://127.0.0.1:9090/unemployment/quarterly')', enable retrieval of the quarterly unemployment data published by the server.

Statelessness:
Each request is independent and includes all necessary information (URL, headers), fulfilling the principle of statelessness.

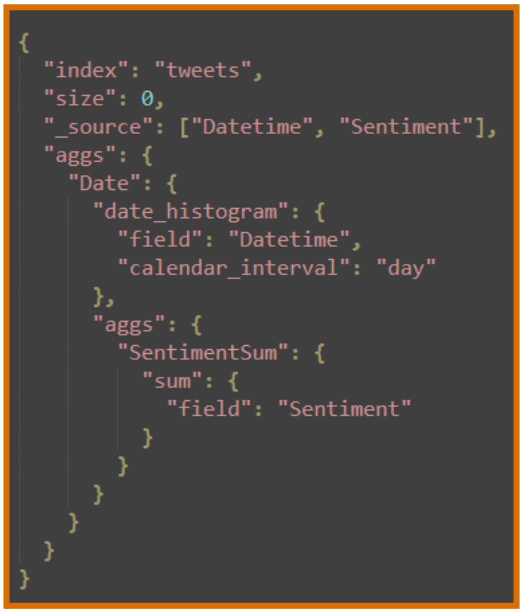
Representation of resources:
All data resources are received in JSON format and converted into Python data frames.

ElasticSearch Queries:



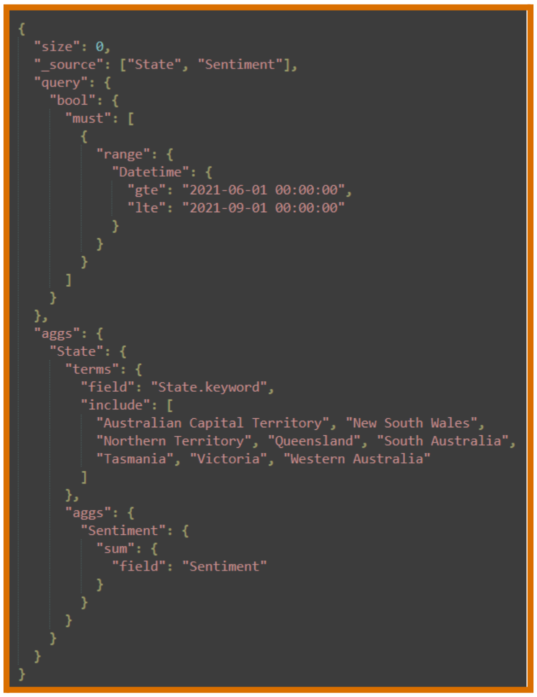
```
{
  "_source": ["Date", "Price", "Change %"],
  "query": {
    "bool": {
      "must": [
        {
          "range": {
            "Date": {
              "gte": "2021-06-21",
              "lte": "2022-06-04"
            }
          }
        }
      ]
    }
  }
}
```

Figure 2.1: An ElasticSearch query that searches the ASX dataset, pulling the date, price ($), and percentage price change (%). It does this for all days between the 21st of June 2021 and the 4th of June 2022.



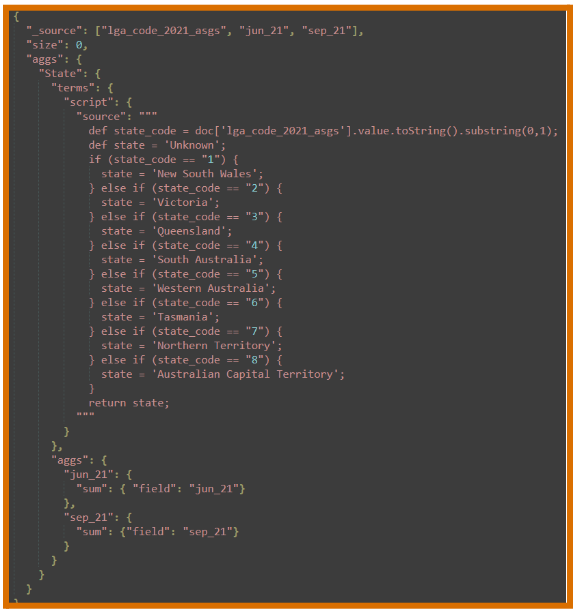
```
{
  "index": "tweets",
  "size": 0,
  "_source": ["Datetime", "Sentiment"],
  "aggs": {
    "Date": {
      "date_histogram": {
        "field": "Datetime",
        "calendar_interval": "day"
      },
      "aggs": {
        "SentimentSum": {
          "sum": {
            "field": "Sentiment"
          }
        }
      }
    }
  }
}
```

Figure 2.2: An elastic search that searches the Twitter dataset and pulls days (in the format of Datetime) and Sentiment. Where Sentiment is the sum of individual Twitter sentiments across that day.



```
{
  "size": 0,
  "_source": ["State", "Sentiment"],
  "query": {
    "bool": {
      "must": [
        {
          "range": {
            "Datetime": {
              "gte": "2021-06-01 00:00:00",
              "lte": "2021-09-01 00:00:00"
            }
          }
        }
      ]
    }
  },
  "aggs": {
    "State": {
      "terms": {
        "field": "State.keyword",
        "include": [
          "Australian Capital Territory", "New South Wales",
          "Northern Territory", "Queensland", "South Australia",
          "Tasmania", "Victoria", "Western Australia"
        ]
      },
      "aggs": {
        "Sentiment": {
          "sum": {
            "field": "Sentiment"
          }
        }
      }
    }
  }
}
```

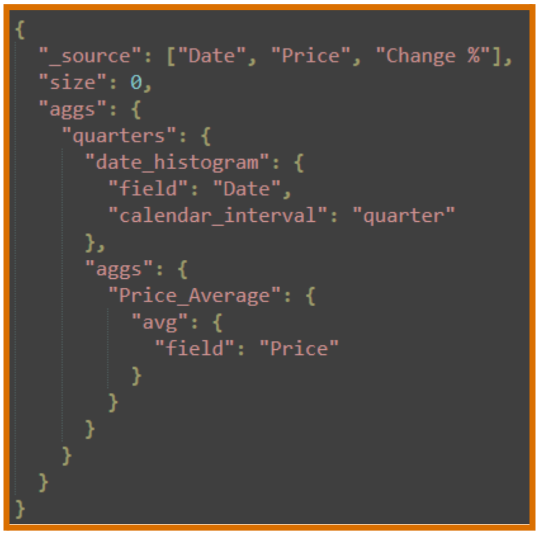
Figure 2.3: An elastic search that searches the Twitter dataset, pulling the sum of Sentiment



```
{
  "_source": ["lga_code_2021_asgs", "jun_21", "sep_21"],
  "size": 0,
  "aggs": {
    "State": {
      "terms": {
        "script": {
          "source": """
            def state_code = doc['lga_code_2021_asgs'].value.toString().substring(0,1);
            def state = 'Unknown';
            if (state_code == "1") {
              state = 'New South Wales';
            } else if (state_code == "2") {
              state = 'Victoria';
            } else if (state_code == "3") {
              state = 'Queensland';
            } else if (state_code == "4") {
              state = 'South Australia';
            } else if (state_code == "5") {
              state = 'Western Australia';
            } else if (state_code == "6") {
              state = 'Tasmania';
            } else if (state_code == "7") {
              state = 'Northern Territory';
            } else if (state_code == "8") {
              state = 'Australian Capital Territory';
            }
            return state;
          """
        }
      },
      "aggs": {
        "jun_21": {
          "sum": { "field": "jun_21"}
        },
        "sep_21": {
          "sum": {"field": "sep_21"}
        }
      }
    }
  }
}
```

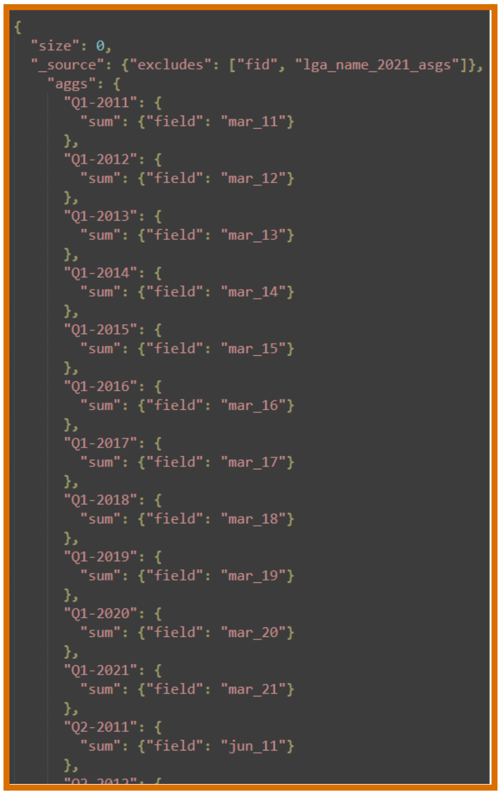
Figure 2.4: An elastic search that searches the unemployment dataset, aggregating and pulling the Australian State, and total unemployment in that State for June and September 2021.

levels across each State. Specifically on the period 06/2021 to 09/2021.



```
{
  "_source": ["Date", "Price", "Change %"],
  "size": 0,
  "aggs": {
    "quarters": {
      "date_histogram": {
        "field": "Date",
        "calendar_interval": "quarter"
      },
      "aggs": {
        "Price_Average": {
          "avg": {
            "field": "Price"
          }
        }
      }
    }
  }
}
```

Figure 2.5: An ElasticSearch query that searches the ASX dataset, pulling the Quarter, Price Average ($) across the quarter.



```
{
  "size": 0,
  "_source": {"excludes": ["fid", "lga_name_2021_asgs"]},
    "aggs": {
      "Q1-2011": {
        "sum": {"field": "mar_11"}
      },
      "Q1-2012": {
        "sum": {"field": "mar_12"}
      },
      "Q1-2013": {
        "sum": {"field": "mar_13"}
      },
      "Q1-2014": {
        "sum": {"field": "mar_14"}
      },
      "Q1-2015": {
        "sum": {"field": "mar_15"}
      },
      "Q1-2016": {
        "sum": {"field": "mar_16"}
      },
      "Q1-2017": {
        "sum": {"field": "mar_17"}
      },
      "Q1-2018": {
        "sum": {"field": "mar_18"}
      },
      "Q1-2019": {
        "sum": {"field": "mar_19"}
      },
      "Q1-2020": {
        "sum": {"field": "mar_20"}
      },
      "Q1-2021": {
        "sum": {"field": "mar_21"}
      },
      "Q2-2011": {
        "sum": {"field": "jun_11"}
      },
      "Q2-2012": {
```

Figure 2.6: An elastic search that searches the unemployment dataset, summing and retrieving total unemployment levels for each quarter from 12/2010 to 9/2010. These are stored in the fields of the data.

Here is a video showcasing our system functionality:
https://www.youtube.com/watch?v=9FhqIM40ksw&ab_channel=Dean

Data Analysis:

Question 1: What is (if any) the statistical relationship between the evaluated sentiments of Australian tweets and the daily ASX 200 from mid-2021 to mid-2022?

To first observe trends of how Twitter Sentiment and the ASX 200 changed with time, the following figures were produced:

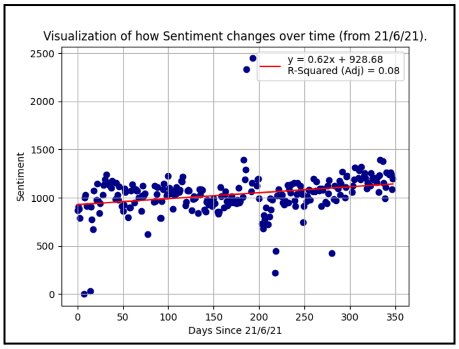


Figure 3.1: A dot plot and linear regression showcasing how daily Twitter Sentiment changes over the time period from 21/06/2021 to 04/06/2022.

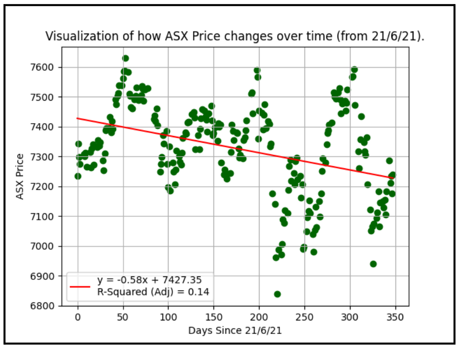


Figure 3.2: A dot plot and linear regression showcasing how the daily ASX 200 closing price ($) changes over the time period from 21/06/2021 to 04/06/2022.

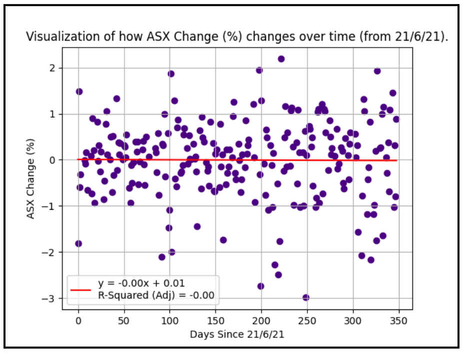


Figure 3.3: A dot plot and linear regression showcasing how the daily percentage change in the ASX 200 Price changes over the time period from 21/06/2021 to 04/06/2022.

All data had a low correlation with the linear model, however, more careful inspections of the shape formed by the dot plots in Figure 3.1, Figure 3.3, and particularly Figure 3.2 suggest that there is an underlying trend and that the linear model may simply be insufficient at detecting it. We will look for correlations between Sentiment vs. ASX Price ($) and Sentiment vs. ASX Change (%) by conducting regressions of one against the other.

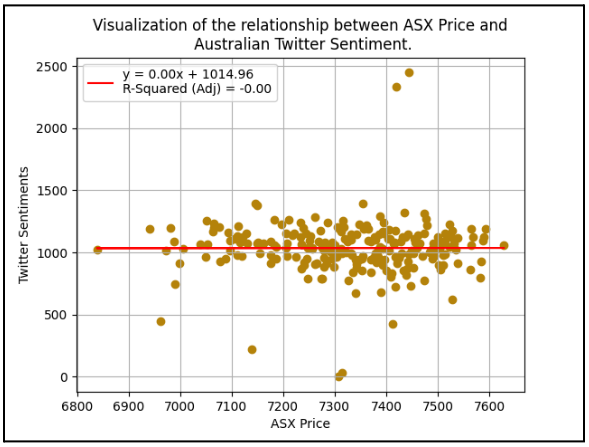


Figure 3.4: A dot plot and linear regression showcasing the correlation between daily ASX 200 Closing Price ($) and overall daily Australian Twitter Sentiments.

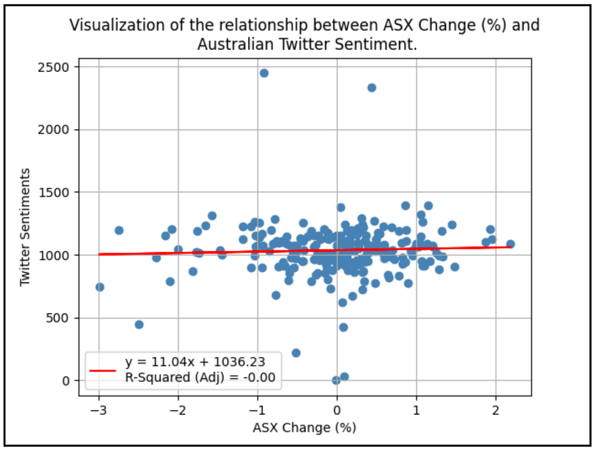


Figure 3.5: A dot plot and linear regression showcasing the correlation between daily percentage change in the ASX 200 Closing Price ($) and overall daily Australian Twitter Sentiments.

Both Figures 4 and 5 clearly indicate any meaningful relationship between ASX Price or Change with Twitter sentiments. The adjusted R-squared of zero on both strongly reinforces this point. This conclusion is further supported by the regression parameter being 0 within ASX price and very slight positive regression in Figure 3.5.

One potential explanation for this lack of correlation is the relatively stable nature of the Australian Stock Exchange. The change from the beginning of the period analysed to the end was only a drop of evaluation of 200 points, making the change a very small 2.7% decrease. This is a very non-volatile change over the period analysed thus increasing the difficulty of finding a noticeable shift in Twitter sentiment.

The events that would most likely display more significant changes in sentiment from ASX changes would involve more significant events. Examples of this would be the recent occurrence of Covid 19, which had a very strong effect on the ASX evaluation. Election results or global issues could also have significant effects on both the ASX and Twitter sentiment, however, the cause of the sentiment change might be more indicative of the event, and less the ASX value dropping.

An important consideration could be that instead of the ASX affecting Twitter sentiment, ASX might be a reflection of broader sentiment. When the public sentiment is increasing or decreased it could have a flow-on effect, increasing or decreasing the ASX valuation. If a positive event occurs, or overall sentiment is up it can lead to an increase in investor confidence and an increase in buying activity leading to a subsequent increase in stock prices, raising the ASX evaluation. Conversely, a trend of negative sentiment could lead to selling off, dropping the evaluation of the market. Therefore, while individual tweets might not directly correlate with ASX changes, the overall mood and sentiment within Australia can have a pronounced impact on the ASX performance. This can be further examined during events that have a far more significant impact upon sentiment, and comparing the drastic change to the ASX to see if there is a correlation.

In conclusion, while the data indicates no meaningful relationships between Twitter sentiments and the ASX, this can be attributed to the stability of the ASX within the analysed period. To draw a more definitive conclusion the question must be framed around events that have greater impacts on public sentiment, to see if the shift in public sentiment will affect the ASX evaluation or the other way around, depending upon which changes first.

Question 2: What is (if any) the statistical relationship between the evaluated sentiments of Australian tweets and the unemployment levels across various states during mid-2021?

Let’s first observe how high the levels of Sentiment and Unemployment are across various Australian States

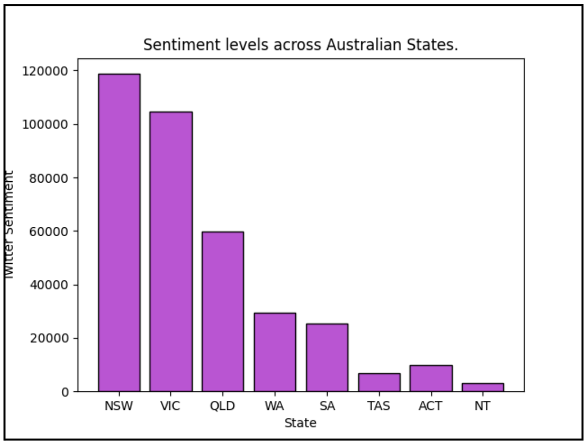


Figure 3.7: A bar plot showcasing the varying levels of total Twitter sentiments across Australian states between 06/2021 and 09/2021.

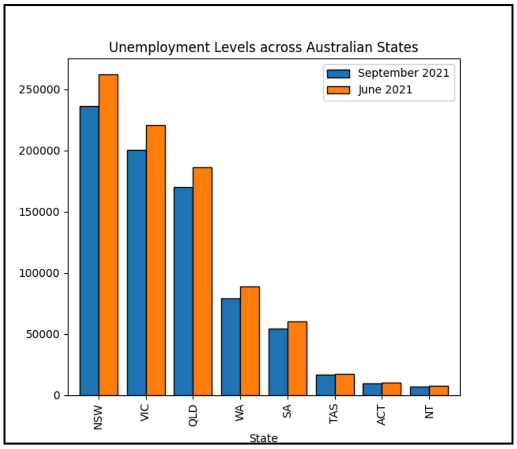


Figure 3.6: A bar plot showcasing the varying levels of unemployment quantities across Australian states for 06/2021 and 09/2021 respectively.

Both plots seem similar in shape, however this appears to also be correlated with the most populous Australian States. Where you would expect higher absolute values of unemployment, and also a greater volume of tweets that would amplify trends. Therefore, we should normalise these quantities by population in order to get a more accurate picture of the question.

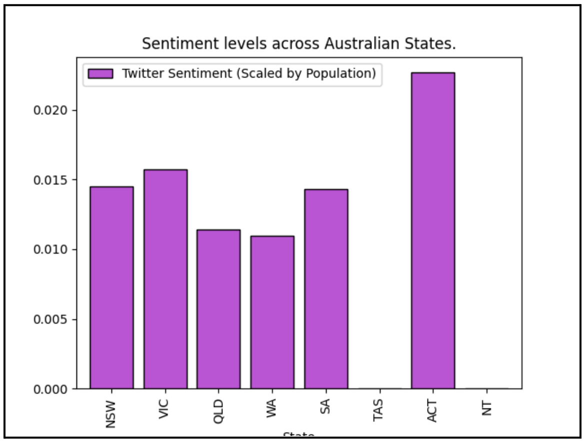


Figure 3.8: A bar plot showcasing the Twitter sentiment scaled (divided) by population for each Australian State between 06/2021 and 09/2021.

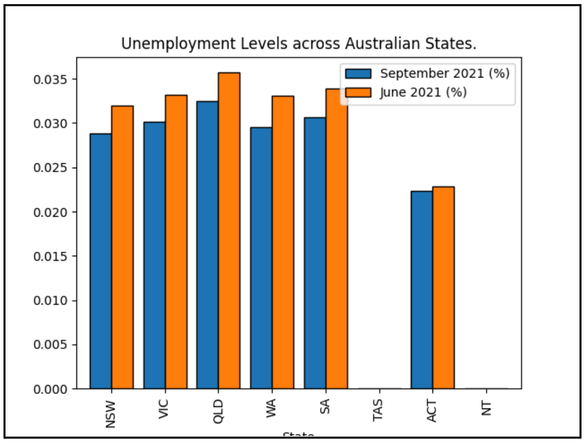


Figure 3.9: A bar plot showcasing the unemployment levels (divided) by population for each Australian State for 06/2021 and 09/2021 respectively.

The shifted data reveals a notable inverse relationship between unemployment and Twitter sentiment across Australia. This is most evident within the ACT (Australian Capital Territory), with the lowest-scaled unemployment and the highest-scaled sentiment. This trend is also visible in other regions such as NSW (New South Wales), VIC (Victoria), QLD (Queensland), and SA (South Australia), forming a convex shape in Figure 3.8 and a concave shape in Figure 3.9. TAS (Tasmania) and NT (Northern Territory) had very minimal data points from the Twitter and unemployment datasets overscaling them, so they were not included within the scaled data analysis.

The inverse relationship between Twitter sentiment and unemployment can be attributed to potential factors. One potential factor is the psychological and economic impact of employment on individuals and communities. When unemployment rates are low, people experience higher levels of financial security and improved mental health. This would have a flow-on effect with more positive expressions and interactions on social media like Twitter. Conversely, as exhibited within the other states, higher unemployment rates exhibit a lower average sentiment. This is because higher levels of unemployment can lead to financial stress, anxiety, and uncertainty, resulting in more negative sentiments being expressed online.

Additionally, lower unemployment can correlate with stronger economic performance. This allows economies and communities to have better job prospects, higher disposable incomes and more consumer confidence. These positive conditions can enhance the overall public sentiment, which would be reflected by twitter average sentiment. This leads to our third question which will look at whether there is a correlation between ASX and unemployment rates.

In conclusion, the inverse relationship between Twitter sentiment and unemployment levels within Australia highlights a strong connection between employment and public mood. Employment status significantly affects an individual's financial and psychological well-being, which in turn would influence the sentiment of the tweets. This relationship highlights the importance of economic stability and job creation in ensuring communities are happy, which can be reflected in social media sentiment.

Question 3: What is (if any) the statistical relationship between the overall unemployment level across Australia and the ASX 200 from 2012 to 2021?

Let’s first observe how high the levels of Sentiment and Unemployment are across various Australian States

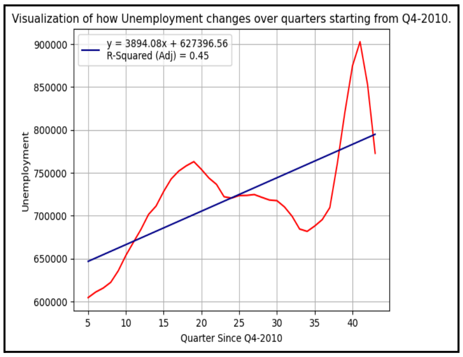


Figure 3.10: A line graph showing a steady and continuous rise in unemployment

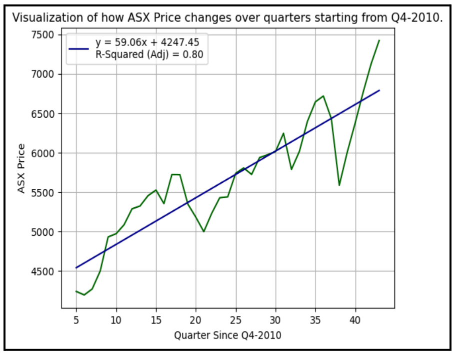


Figure 3.11: A line graph showing a steady and continuous rise in ASX price

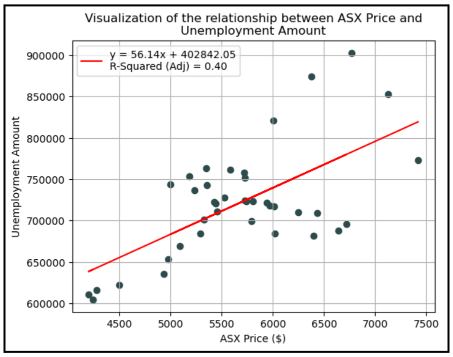


Figure 3.12: A scatter plot displaying the relationship between ASX price and unemployment volume

These graphs were adjusted because the constant upwards trend in unemployment can directly be linked to the increase in population. The ratio of unemployed to employed could remain constant through the 10 year timeframe, however it is hard to see that if we only look at the raw values, instead of looking at the percentage. Of unemployment

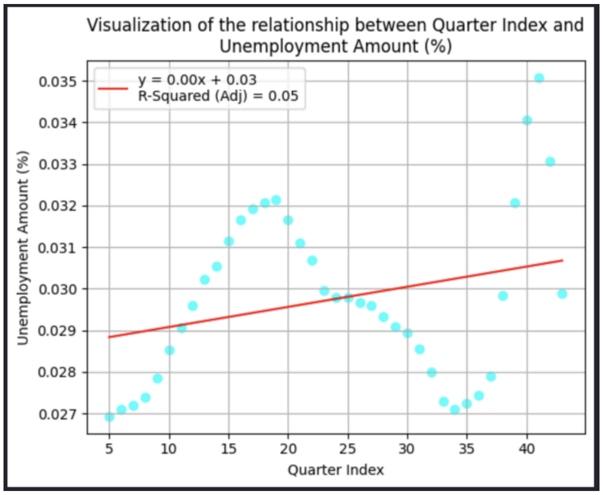


Figure 3.13: A dot plot displaying the portion of the population that is unemployed nationwide in Australia

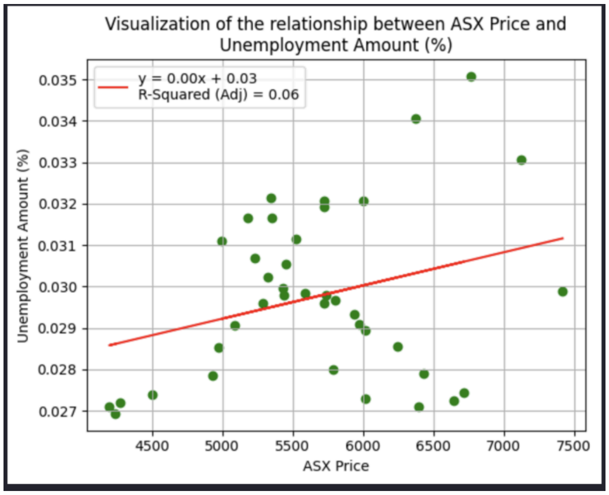


Figure 3.14: A dot plot displaying the relationship between ASX price and unemployment percentage

To determine the statistical relationship between the overall unemployment levels across Australia and the ASX between 2011 and 2021 we analysed the unemployment trends, converting it from numbers to percentage of population, alongside stock market performance. The unemployment data revealed a gradual increase from 2.7% to 3.2% by 2015 and then slowly decreasing back down to 2.7% by 2019. The unemployment rate then spikes to 3.5%, which was in the midst of covid. The upwards trend of the data is not necessarily indicating a higher or lower average employment rate over the 10 year segment, purely due to the fact of a significant economic disruptor distorting the data, being the COVID-19 pandemic.

In contrast the ASX 200 index exhibited an incredibly sporadic pattern with no clear correlation between stock price and unemployment. As seen around the stock evaluation of around 6400, unemployment rates were found at 3.4% and 2.7%, showing no real correlation between the two.

The R-Squared value shows that there is a very weak correlation in both figures. First, the stock market is influenced by a multitude of factors, not limited to the unemployment rate. These additional factors all contribute to the resulting change in ASX price. That is not to say that unemployment rate doesn’t affect the ASX price, instead it in conjunction with a multitude of other variables affect the ASX pricing.

Despite the weak correlation identified in the analysis, it provides a strong basis for understanding the nuanced relationship between unemployment and stock market performance. Future research could benefit from incorporating a larger number of variables that potentially influence ASX prices, thereby determining the strength of each variable's impact in relation to others.

While there appears to be a slight positive relationship, it can be reasonably assumed that this is primarily due to the unique economic event of COVID-19, rather than unemployment playing a key role. This finding emphasises the complexity of financial markets, highlighting the necessity of considering a broader range of variables. It's not one individual factor that affects the market, but rather a culmination of multiple factors.

Discussion and Implications

The comprehensive analysis of the relationship between the ASX 200, unemployment levels from 2011 to 2021 and the twitter sentiment from 2021 to 2022 provided valuable insights into the interdependence of these variables. Despite the weak correlations found, the study has shown the complexity of market performance and public sentiment.

Twitter Sentiment and ASX 200 (2021-2022)

The investigation into the relationship between these two datasets revealed no significant correlation. The adjusted R-squared values were close to zero, indicating that daily fluctuations in Twitter sentiment were not indicative of changes in the ASX 200 index. This lack of correlation can be attributed to multiple factors. The ASX being incredibly stable within this period, only experiencing an overall change of 2%. Also, the nature of tweets not always correlating to ASX fluctuations. Future research upon this topic could involve directly searching for tweets relating to keywords such as: stock, ASX and other stop related topics, however this might have a greater level of distortion, because their sentiment is not really indicative of the overall sentiment of Australians in relation to ASX fluctuations.

Twitter sentiment and Unemployment Levels (2021)

The investigation into the relationship between these two datasets did reveal an inverse relationship. Regions with lower unemployment rates, such as ACT, displayed higher average twitter sentiment scores.The states that had higher unemployment rates shared a lower average sentiment score, indicating an inverse relationship between employment rate and twitter sentiment. These findings enforce the importance of economic stability and job creation in maintaining a positive public mood.

Unemployment levels and ASX 200 in quarters (2011-2021)

The investigation into the relationship between these two datasets did not reveal a significant correlation. Despite both unemployment and ASX displaying overall trends, there was little to no relationship. This suggests that while unemployment may affect the ASX, it is not the only factor, and it is in conjunction with a multitude of other factors. Instead unique economic events are more likely to affect the ASX, and would be displayed in both data fields, as seen with ASX and unemployment both encountering fluctuations at the beginning of COVID-19.

Implications and Future Research

Overall the findings emphasise the complexity of financial markets and the need to consider a mix of variables when analysing market fluctuations. Individual factors such as twitter sentiment and unemployment have some impact on market performance, these are not all the factors that affect the market. Future research should incorporate a larger number of variables, and search as to whether one variable impacts the market more than the others. Additionally, focussing on periods of higher volatility or narrowing down on a specific event and analysing sentiment shift and market fluctuations might provide more insightful results regarding the interactions between sentiment, unemployment and market performance.

Links:

GitHub Repository: https://github.com/TianzTong/comp90024group75
System Functionality Demo:
https://www.youtube.com/watch?v=9FhqIM40ksw&ab_channel=Dean