COMP90024 - Cluster and Cloud Computing

University of Melbourne

Semester 1, 2021

Assignment 2

City Analytics on the Cloud

An investigation of political inclination and various factors in Greater Melbourne

Group 53 members:

Name	**Student ID**	**City**
Aditi Basu	1178282	Perth, Australia
Kevin Van	995203	Melbourne, Australia
Linan Jia	806003	Melbourne, Australia
Nand Lal Mishra	1245159	Uttar Pradesh, India
Zhirui Liang	1255971	Melbourne, Australia

Table of Contents

Introduction

This project examines the characteristics of the Greater Melbourne area by harvesting tweets originating from this area and performing data analytics on them to extract insights about specific aspects, discussed in the next section. This analysis was performed by implementing a cloud-based application that leverages No-SQL document databases like CouchDB, analytical tools like MapReduce and TextBlob, and front-end tools such as PowerBI and Nginx, Docker for containerized deployment and Ansible for dynamic deployment .

Functionality / Scenarios

The scenarios tested in this project are:

1) Do certain Statistical Area 3 (SA3) regions have certain political inclinations? Does a certain political inclination correlate with the crime rate in that region?
2) Which political party is most associated with a negative sentiment?
3) Are SA3 regions with negative sentiment scores more prone to higher crime rates?
4) Are people happier who live in areas that have the highest housing prices?

These scenarios were chosen considering what would be most relevant to the average civilian, such as crime rate, housing prices and how different political affiliations affect our daily lives. At the time of choosing these scenarios, it was also taken into consideration how easy it would be to address these scenarios with the data available to us.

In these scenarios, the political inclination of a SA3 region is denoted by their preference of either the Liberal Party of Australia (engaging in right-wing politics) or the Australian Labor Party (engaging in left-wing politics). Sentiment scores of a SA3 region are evaluated by extracting the sentiment scores of tweets originating from that region and averaging them. Housing prices of a SA3 region are denoted by the region’s median housing price.

For the purposes of this project and proof of concept, the above scenarios have been limited to the Greater Melbourne area, but these scenarios can be extended to other capital cities in Australia.

User Guide

Important:

- Do not delete cluster-1 instance on MRC. A static IP (172.26.131.188) is required for PowerBI so we must keep this one machine up. Otherwise we need to change the IP manually on PowerBI.
- You must have Ansible installed on your machine

Dynamic Deployment with Ansible

1. Change directory to ansible/
2. Replace the “openrc.sh” with your own and replace the line with the “openrc.sh” in “up.sh”
3. Place the private key in ~/.ssh/ and change the key variables in /ansible/host_vars/mrc.yaml and /ansible/host_vars/applications.yaml
4. Run up.sh and input required details as prompted (number of instances, openrc password and sudo password).

Notes

- This deployment only scales up. Scaling down is not properly supported and may cause some issues if you input a number lower than the current number of machines deployed. To properly scale down, CouchDB shards have to be transferred from the machines that are to be shut down and then the machines removed from the cluster. Any other system applications on removed machines must also be started on existing nodes. The user must be connected to the UniMelb network.

Troubleshooting

If you come across any 503 errors please re-run the script or specific component of the script again because there are probably some network issues with the MRC.

Website access

Website IP address is outputted at the end of the up.sh script. (cluster-1 ip address). However, if the static IP address mentioned above (172.26.131.188) is reserved, then this IP address can be used to access the website.

Fauxton Access

Via any ip in ./inventory/hosts.ini at port 5984 and /_utils

System Architecture and Design

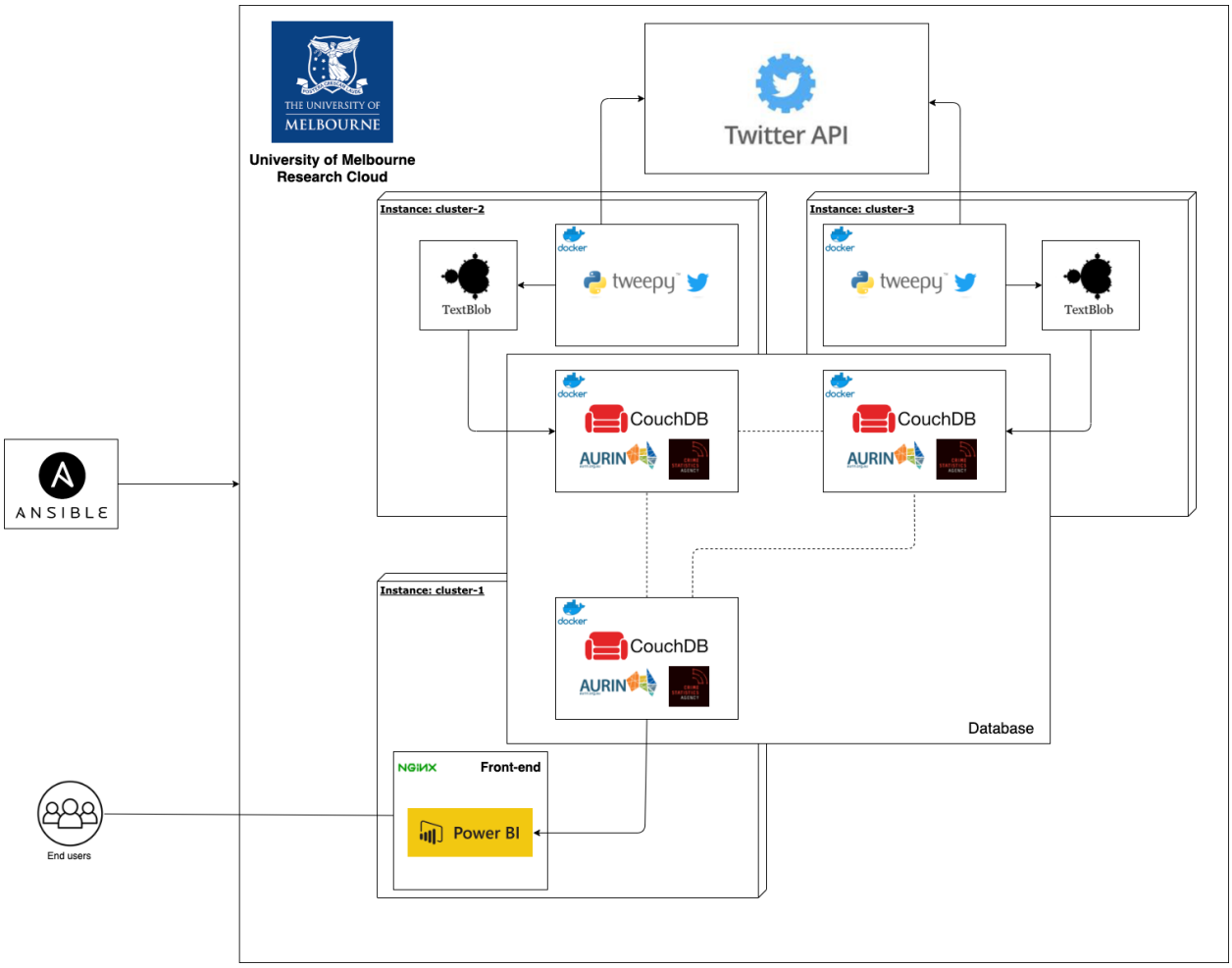

Figure 1: System Architecture

Ansible allowed quick deployment and redeployment from scripts. Each time we need to scale up or down, we do not need to interact with the Melbourne Research Cloud dashboard to create/delete instances, security groups, volumes etc. We can just run one script that we wrote called “up.sh” and specify how many instances we want. A con would be researching and learning the different modules required to set up instances and interact with the instances in order to write the script. This took a long time and was difficult. The resulting advantage is that we can replicate this dynamic deployment when we deploy new applications on the open cloud much more quickly now.

Docker was used for deploying CouchDB, this allowed us to easily run CouchDB without installing its prerequisites on the actual machine, simplifying the process. A con for running it in a Docker container is that all data is lost when the container is stopped. To counter this, we mounted a folder from outside the container into the container to prevent data loss from container crashes or restarts. Since we solved the major con of data loss, we decided to stick with Docker for CouchDB.

Harvester

A Twitter harvester is created by combining Twitter’s Search API and Filtered Stream API with Tweepy. Tweepy is an open source python library that enables accessing the Twitter API. To use Tweepy, one must have a Twitter developer account.

Twitter provides two different API’s for accessing its data. With the Search API, users can pass certain keywords (called tracks) to get tweets that contain these keywords from the past. For regular developer account holders, the search limit is 7 days, (i.e. users will only get tweets with the keywords from the past seven days). In order to increase this limit, one can apply for an academic researcher developer account. The filtered Stream API lets users access tweets as they come through (i.e. users have access to live tweets). To use the filtered Stream, users should provide a list of keywords and will get all the live tweets with those keywords. However, Twitter doesn’t allow the use of keyword filtering and location filtering at the same time.

There are additional limitations for regular developer account holders such as 500,000 tweets/month per project, 450 requests/ 15 min PER app auth. All the tweets harvested from Twitter are stored on the CouchDB cluster on MRC.

DB

CouchDB is highly suited to a distributed environment. A con was that it was very difficult to get a cluster up and running on the remote instances as indicated by the large number of questions on the discussion board about setting up a database cluster. The documentation was also dense and unhelpful/misleading at times. However, after setting up the database, we saw that it was quite useful with its built-in MapReduce functionality. This allowed us to extract useful bits of data (views) to support our analysis scenarios. Due to the time invested in testing out CouchDB, we didn't have much time to test out alternatives and because it was highly recommended by the teaching team, we decided to use CouchDB.

Front End

The website has been developed using HTML, CSS and JS. The web page style is provided by "templatemag", which is a free open source Bootstrap template website, and allows PowerBI data visualisations to be embedded into the website. We also use Nginx as our web server (as opposed to an Apache server), as Nginx uses fewer resources and can be used with Ansible, which can achieve rapid deployment. The home page is shown below.

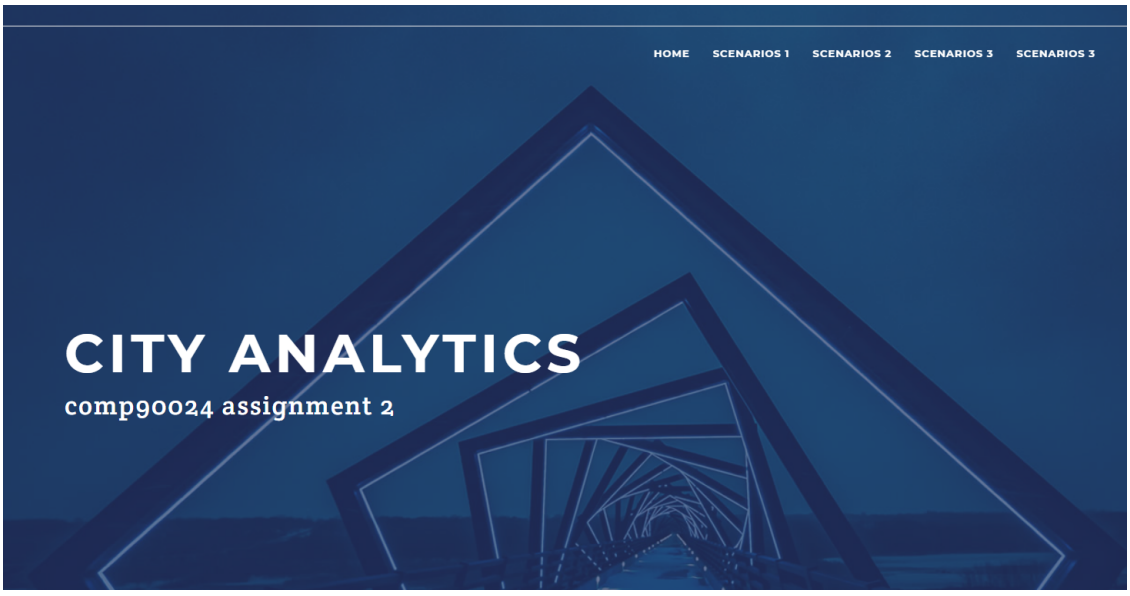

Figure 2: Website homepage

PowerBI

PowerBI is a business analytics service by Microsoft. It aims to provide interactive visualizations and business intelligence capabilities with a simple interface enough for end users to create reports and dashboards.

Our group decided to use PowerBI as part of our frontend solution to utilize its visualization functionality provided by Power BI.

In order for our team to collaborate together on building the dashboard, PowerBI pro license is required, so we have signed up for an Office 365 E5 trial account, which includes the PowerBI Pro license that is needed to publish the content to the PowerBI service.

Data Modeling

Data Modeling is one of the features in PowerBI to connect multiple data sources through a relationship. A relationship defines how data sources are connected with each other and you can then create data visualizations on multiple data sources.

With the modeling feature, we can build custom calculations on the existing tables and these columns can be presented as visualizations. This allows us to define new metrics and to perform custom calculations for those metrics.

The screenshot below is the data model we designed for our group project. The SA3 regions table at the top is the dimensional table, which is used to link with all other fact tables used in different scenarios.

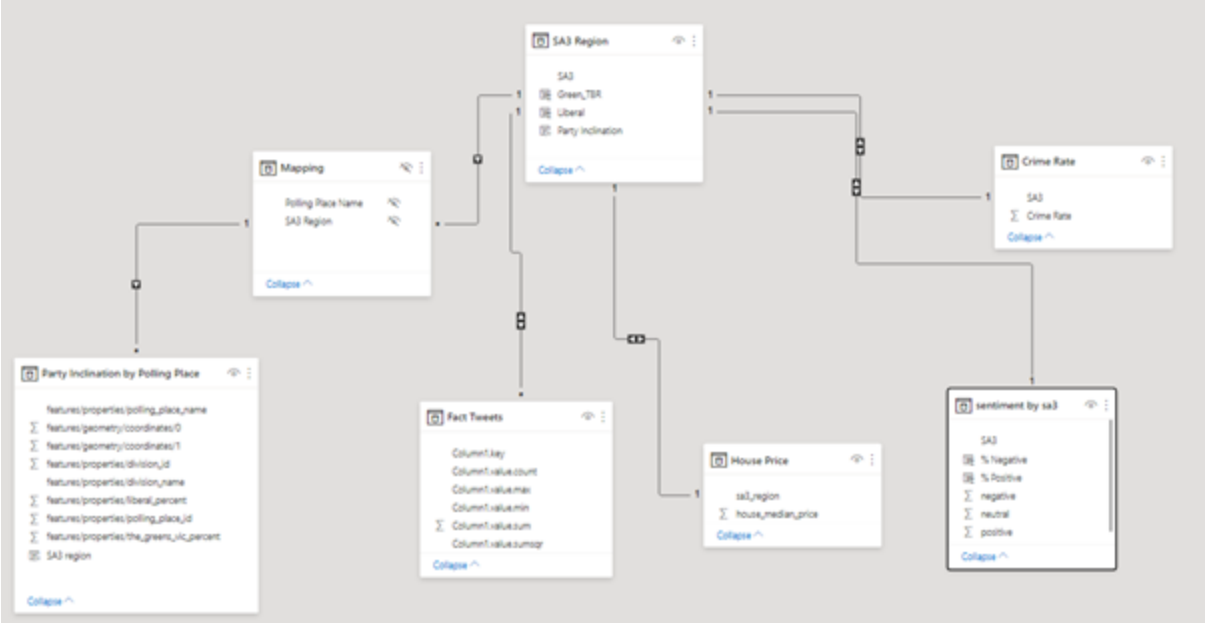

Figure 3: Data model on PowerBI

Refreshing Data

Refreshing data means importing data from the original data sources into a dataset, either based on a refresh schedule or on-demand. In PowerBI service, we set the number of daily refreshes to 8, which is necessary since the underlying Twitter data in CouchDB changes frequently.

The screenshot below shows the configuration set in the PowerBI service.

Figure 4: Configuration for schedule refresh in PowerBI

Data Gateway

The on-premises data gateway acts as a bridge to provide a secure data transfer between on-premises data (data that resides on MRC) and some Microsoft cloud services. By using a gateway, we keep CouchDB databases on its on-premises network, yet securely access that data in PowerBI services.

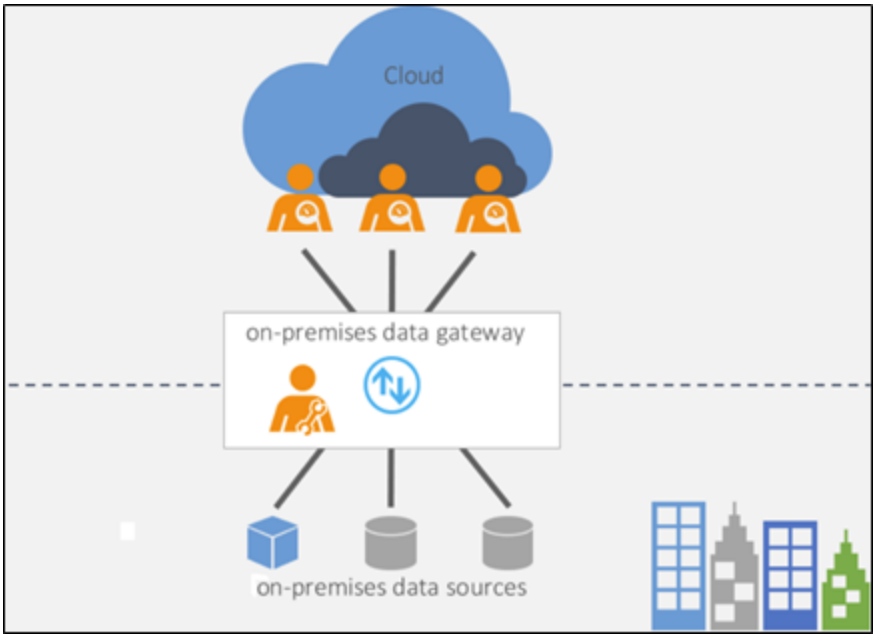

Figure 5: On-premises data gateway on PowerBI

A windows OS is needed here in order to get the gateway installed. Currently, we are running Windows VM with VPN connected on our laptop. Future work would include getting a windows server machine up and running on the MRC.

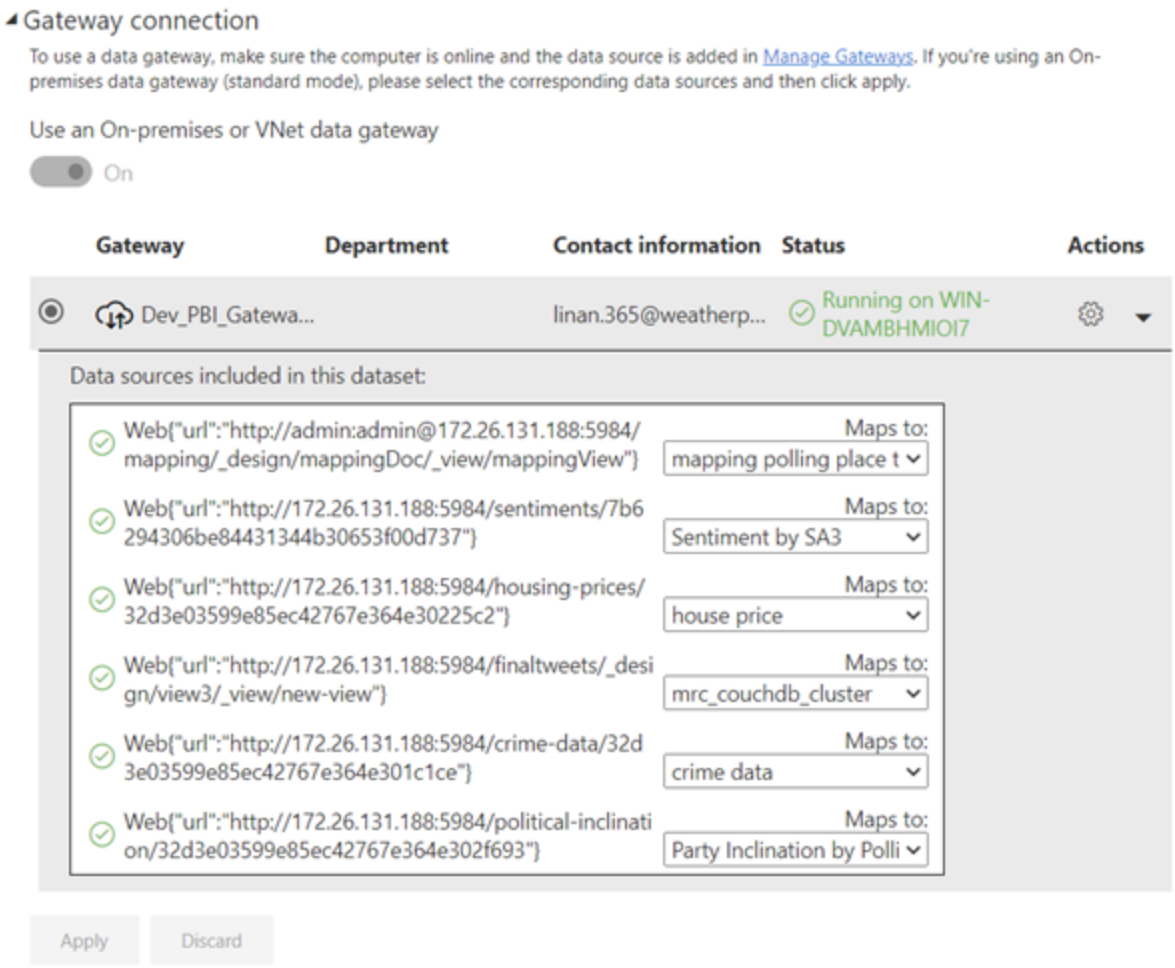

Figure 6: Gateway connection on PowerBI

Pros and Cons of the Unimelb Research Cloud

The University of Melbourne Research Cloud allowed us to build a highly specific distributed Twitter analytics website that could scale according to our needs. However, building such an application took a very long time for our team and had a very high learning curve to overcome where we needed to. As a result not as many people not as many people use the Melbourne Research Cloud compared to Azure and AWS. In comparison, services like AWS and Azure offer many functionalities such as the scaling that we had to implement that the developer can utilize at a click of the button. This allows developers that do not want to suffer through a very steep learning curve to deploy applications and/or for those who just want to save time.

Twitter Analytics

The analysis of tweets has been done using libraries such as TextBlob and shapley. TextBlob is a simplified text processing library for python. TextBlob is built on top of the NLTK (Natural Language Toolkit) library, which is another text processing library for python, but is very extensive, and TextBlob provides a simplified version of it that was simpler to use for the purpose of our project.

Another important library used for Twitter analytics is shapely. Shapely package for python enables the manipulation and analysis of planar geometric objects. We have used various modules such as Polygon, box, Point of shapely in this project. Major use of this library is for classification of tweets into different SA3 regions. The bounding boxes of each SA3 region and the bounding box of the tweets are converted into Polygons and then we check if the centroid of the polygon of the tweet lies in any of the polygons of the SA3 regions. If it does, we classify the tweet as belonging to this region, else we classify it as ‘RestOfAus’ and this is not needed for analysis.

MapReduce proved to be very useful in performing our analysis as it allowed a large number of documents in the database to be processed very quickly. One of the ways we used MapReduce on our tweets was to use the Map function to specify the SA3 region as the key, and “1” as the value. This allowed us to then use the Reduce function to sum the values per SA3 region and thus find the total number of tweets in each region and utilise these numbers in our analysis. There were also other use cases of MapReduce in our project, and it is further explained in our YouTube video (refer to Helpful Links).

Discussion and Analysis of Scenarios & Data

Data Description

The data for this project was gathered from a variety of sources, including the Twitter API, AURIN, Crime Statistics Agency Victoria, and from the bigTwitter.json file made available to us for Assignment 1. Because the tweets from the Twitter API only contained bounding box coordinates to describe their location, the bigTwitter.json file was deemed as a good alternative

since it contains over 200,000 geo-located tweets. These tweets were then analysed by performing the following steps:

1. Two lists (within keywords.csv) were created that consist of words, hashtags and phrases that could potentially indicate that a tweet contains political content. One list contains keywords typically associated with the Australian Labor Party (or left-wing Australian politics more generally), and the other contains keywords typically associated with the Liberal Party of Australia (or right-wing Australian politics more generally).
2. Tweets from bigTwitter.json were filtered such that our final tweet collection contained tweets from only the Greater Melbourne area. Around 400k tweets were collected from this area and classified into SA3 regions based on the geocodes.
3. The percentage distribution of tweets in the SA3 regions in these 400k tweets was studied and then 15k tweets from the 400k tweets were selected so as to maintain the percentage distribution of tweets in the SA3 regions. All this processing was done on UniMelb's high performance computing system, Spartan.
4. A second filter was applied on tweets from bigTwitter.json using these lists (i.e. if the tweet contains one or more of these keywords, it will be extracted from the corpus and classified as containing political content. From the tweets extracted in Step 2, ~80 tweets contained political content but none of the tweets were definitive in their inclination towards any political party.
5. Using TextBlob, a Python library for processing textual data, the sentiment 'score' for the tweets obtained from Step 2 were evaluated. Each tweet was classified as either positive, negative or neutral.

AURIN was used to gather data related to political party preferences of the various SA3 regions in the Greater Melbourne area. To evaluate the political inclination of a SA3 region, a dataset was first extracted from AURIN, containing polling place names, and their corresponding percentage of votes for the Liberal Party. For reasons unknown, there were lots of missing data points for the percentage of votes won by the Labor Party, and hence this data was not included in the dataset extracted from AURIN. Instead, a region was classified as right-winged if the percentage of votes won by the Liberal Party was >35%, and left-winged otherwise. This was chosen as the threshold because the Nationals party (another right-winged party) is in a coalition with the Liberal Party, and a buffer of about 15% (reaching 50% of right-winged votes) would allow consideration of these votes for the Nationals rather than solely looking at Liberal Party votes. This dataset is from the year 2019, as current data wasn't available on AURIN.

While the original intention was to extract political inclination of regions from solely the tweets, this task turned out to be unfeasible because only 80 tweets out of the original >200,000 tweets contained some form of political content, with none of these tweets exhibiting any inclination towards a political party. AURIN was thus used to gauge political inclination of SA3 regions.

AURIN was also used to gather data about housing prices in the Greater Melbourne area. Housing prices were reflected by the median housing price of houses (as opposed to apartments) in each SA3 region. Due to lack of current data, the most recent dataset, from the year 2017, was chosen instead.

Data related to crime rates in the SA3 regions, in the year 2020, were obtained from Crime Statistics Agency Victoria.

Results

Note: All visuals included in this section have been developed on Power BI and embedded in our project website. To experience full functionality of the figures shown, please visit our website.

Scenario 1

Do certain Statistical Area 3 (SA3) regions have certain political inclinations? Does a certain political inclination correlate with the crime rate in that region?

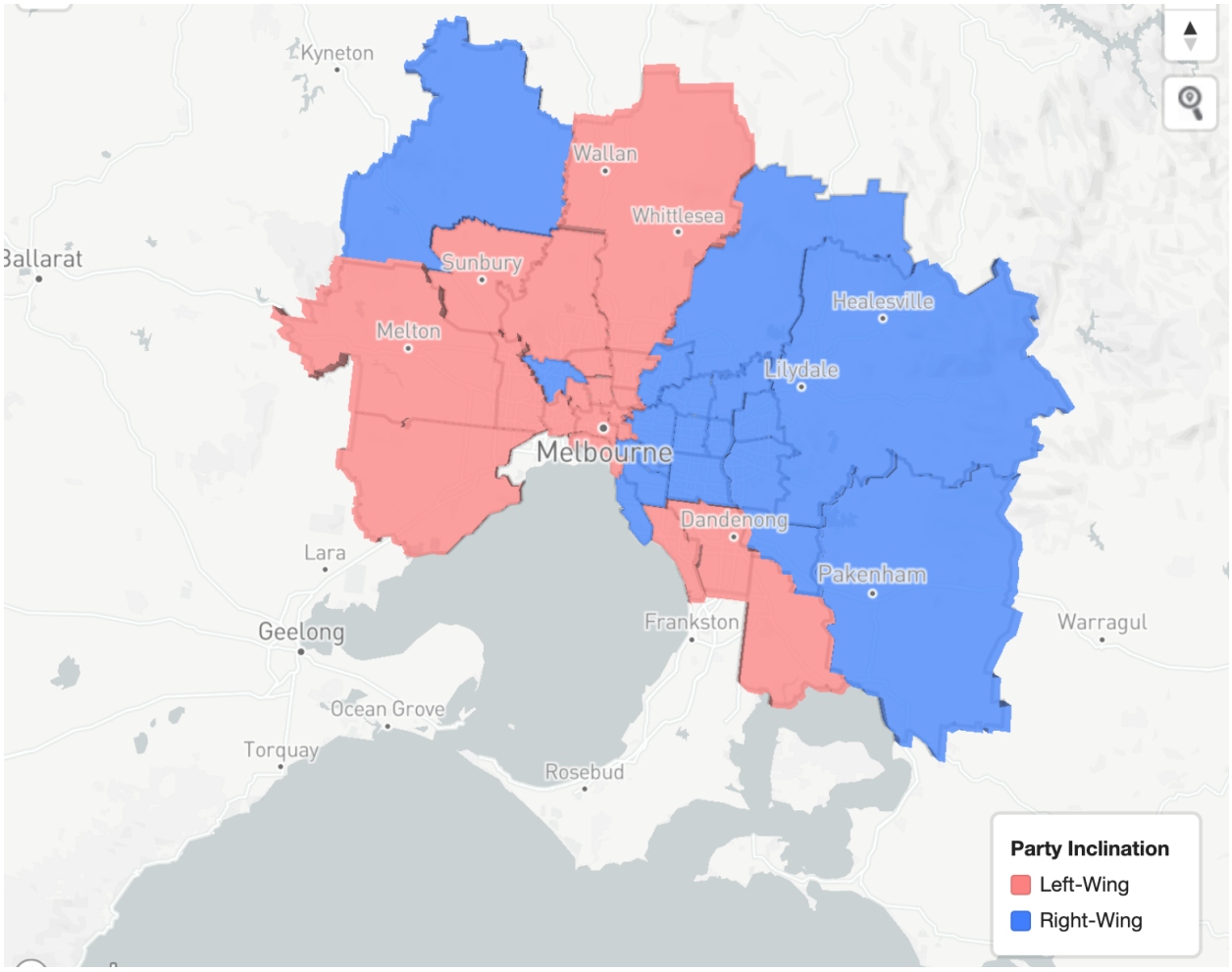

Figure 7: Map of greater Melbourne SA3's with the heights being crime rate

The figure above (Figure 7) shows the distribution of political inclination (either left-wing or right-wing) among SA3 regions in the Greater Melbourne area. Additionally, the project website also portrays the crime rate in each SA3 region using the region's height in the map.

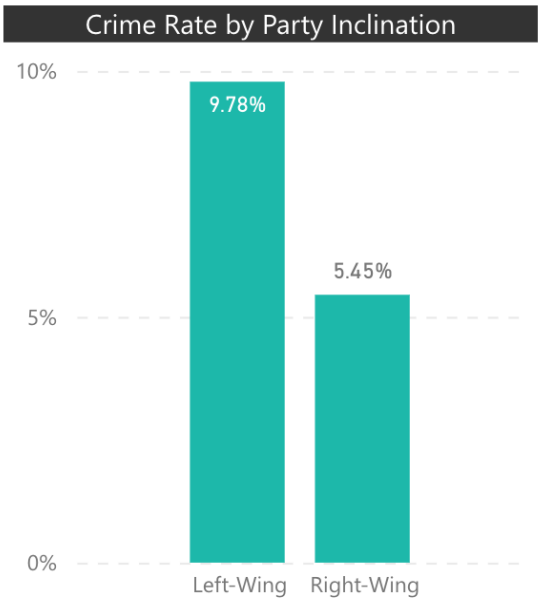

Figure 8: Bar chart of crime rate per party inclination

The figure above (Figure 8) explicitly shows the average crime rate for left-wing regions and right-winged regions. From a first glance, it seems that left-wing regions tend to have higher crime rates. While this correlation might be true, it may not necessarily translate to causation (i.e. supporting left-wing politics does not cause people to commit crimes). There are a variety of factors within a city or a region that affect the local crime rate, and these may not even include general political party affiliation in that region. For the user's ease of understanding, the figure below (Figure 9) depicts the crime rate in a few of the SA3 regions.

Party Inclination	SA3	# Tweets Collected	Crime Rate
Left-Wing	Melbourne City	1122	19.48%
	Yarra	1204	13.79%
	Port Phillip	483	11.55%
	Dandenong	383	11.40%
	Brimbank	92	9.63%
	Maribyrnong	466	9.57%
	Darebin - North	411	9.50%
	Darebin - South	417	9.50%
	Melton - Bacchus Marsh	68	8.36%
	Moreland - North	356	7.31%
	Whittlesea - Wallan	414	7.09%
	Kingston	525	6.97%
	Wyndham	632	6.45%
	Casey - South	29	6.33%
	Brunswick - Coburg	413	

Figure 9: Crime rate in Greater Melbourne SA3's

Scenario 2

Which political party is most associated with a negative sentiment?

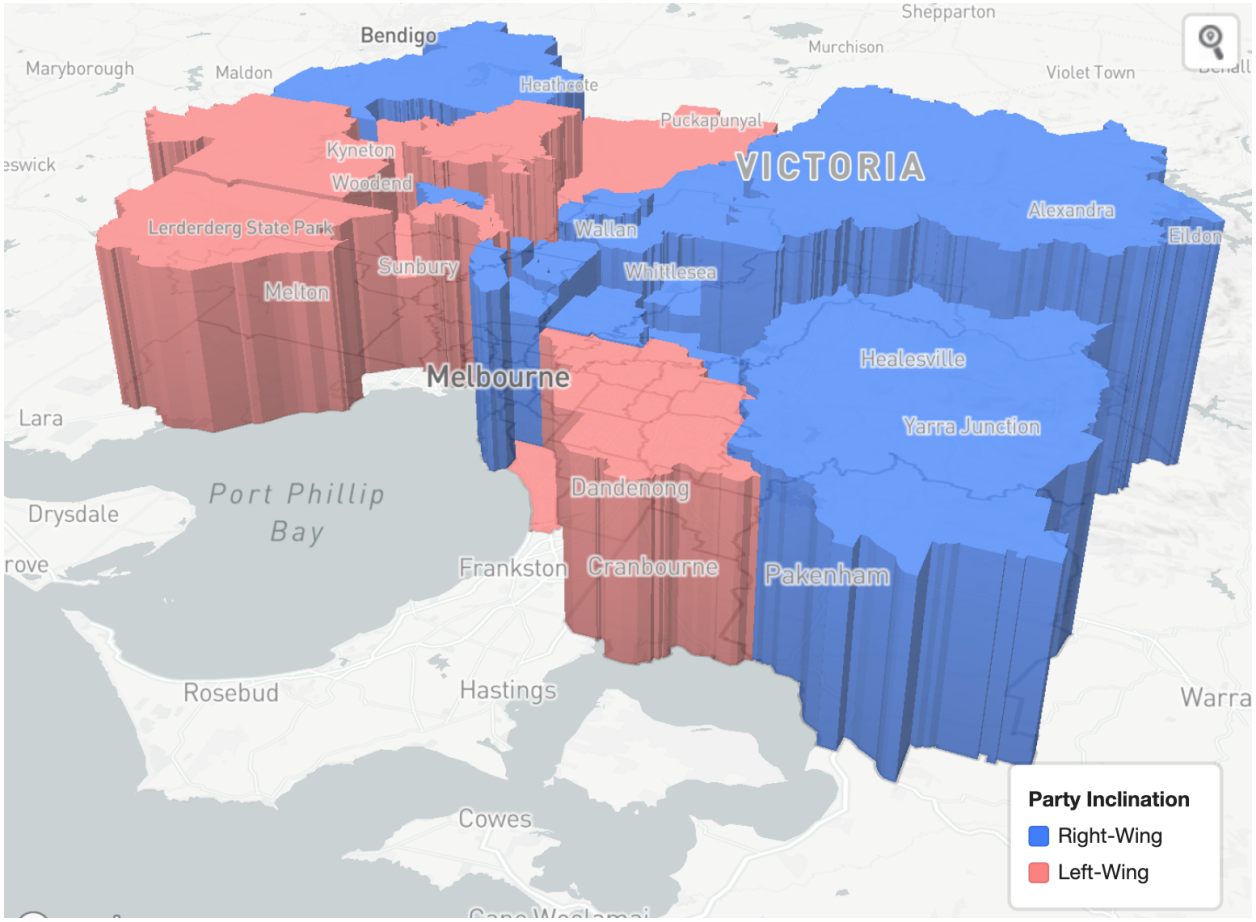

Figure 10: Map of party inclination and Positive Sentiment

The figure above (Figure 10) shows, for each SA3 region, its political inclination (by colour in the legend) and its general sentiment (by height). The higher the region, the more positive the sentiment. The bar graph in the figure below shows the percentage of positive and negative sentiments for each political party, and it can be observed that they approximately display similar percentages of positive and negative sentiments. The pie chart on the left can be used to depict the sentiment distribution of each region without taking political affiliation into consideration.

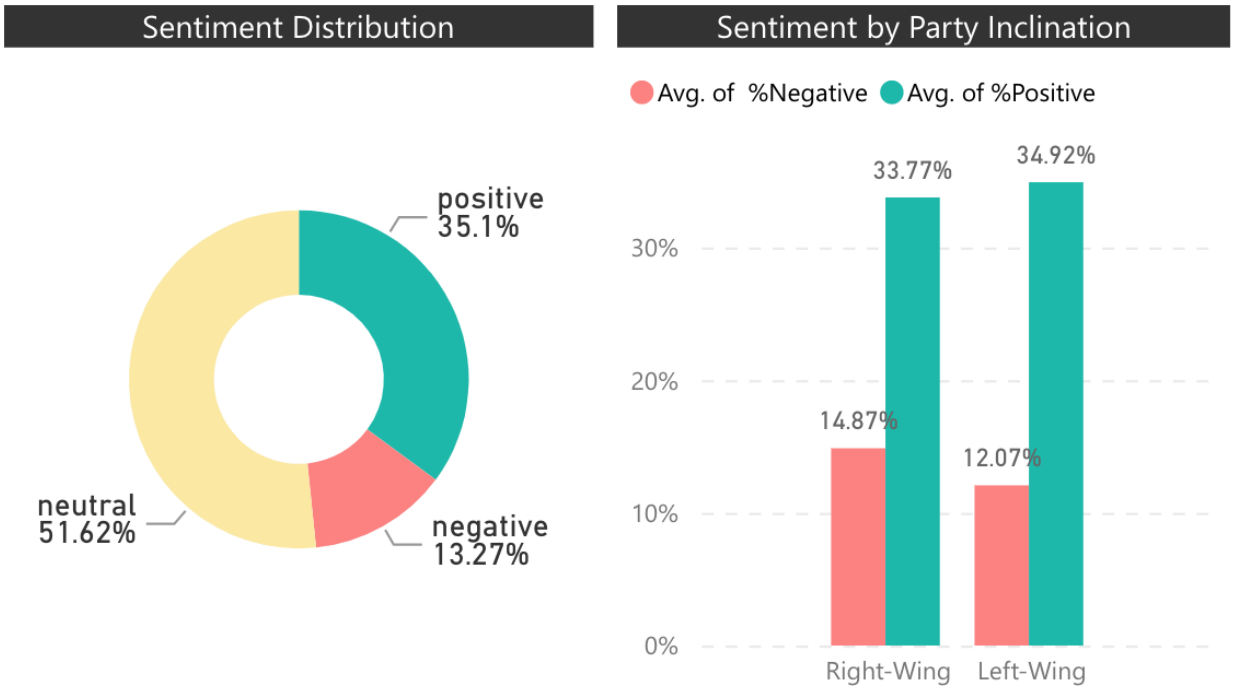

Figure 11: Pie chart and Bar chart of sentiment and party inclination vs sentiment

Party Inclination	SA3	# Tweets Collected	% Positive Tweets
Left-Wing	Yarra	1204	51.29%
	Melbourne City	1122	39.16%
	Wyndham	632	39.30%
	Kingston	525	4.96%
	Port Phillip	483	42.12%
	Maribyrnong	466	38.06%
	Darebin - South	417	36.78%
	Whittlesea - Wallan	414	27.12%
	Brunswick - Coburg	413	38.11%
	Darebin - North	411	34.39%
	Essendon	409	33.33%
	Dandenong	383	29.06%
	Moreland - North	356	31.83%
	Hobson Bay	280	37.99%
	Sunbury	127	33.33%

Figure 12: List of tweets per region and % positive tweets

Scenario 3

Are SA3 regions with negative sentiment scores more prone to higher crime rates?

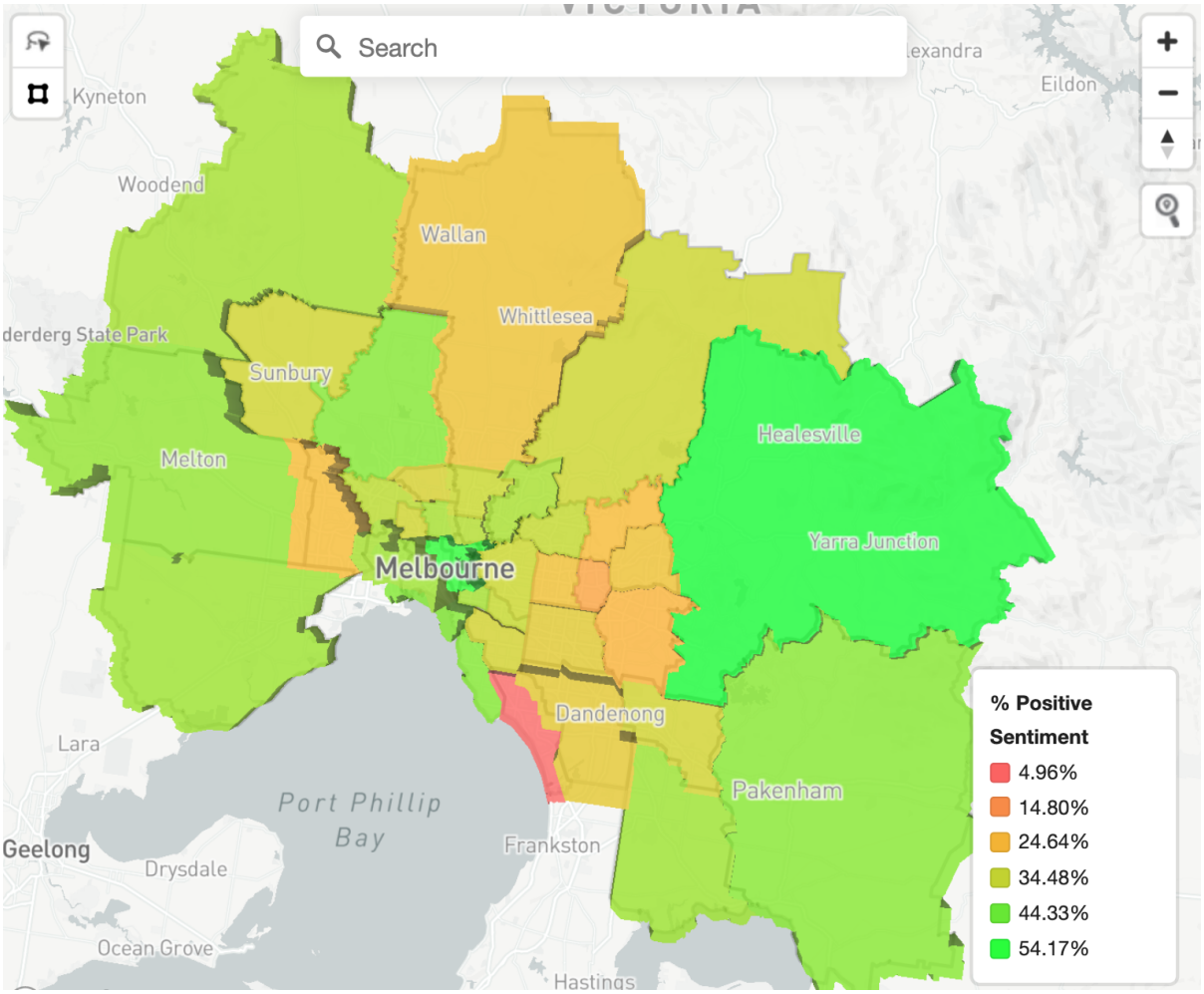

Figure 13: Height Map of positive sentiment in SA3 regions

The figure above (Figure 13) shows the proportion of tweets with positive sentiments in each region. Coupling this figure with the scatter plot below, it is observed that there exists a positive correlation between positive sentiment and crime rate (i.e. the more positive the sentiment, the higher the crime rate). This correlation obviously goes strongly against human intuition, and can be explained by the fact that the sentiment for each region has been calculated by solely evaluating the sentiment score of tweets from that region. People who tweet make up only part, and not all, of the population present in a region. It can be argued that people who do not spend time on social media tend to be happier than people who do (Robinson, 2021). As such, using

tweets to evaluate the sentiment of a region may actually lead us away from what the reality actually is.

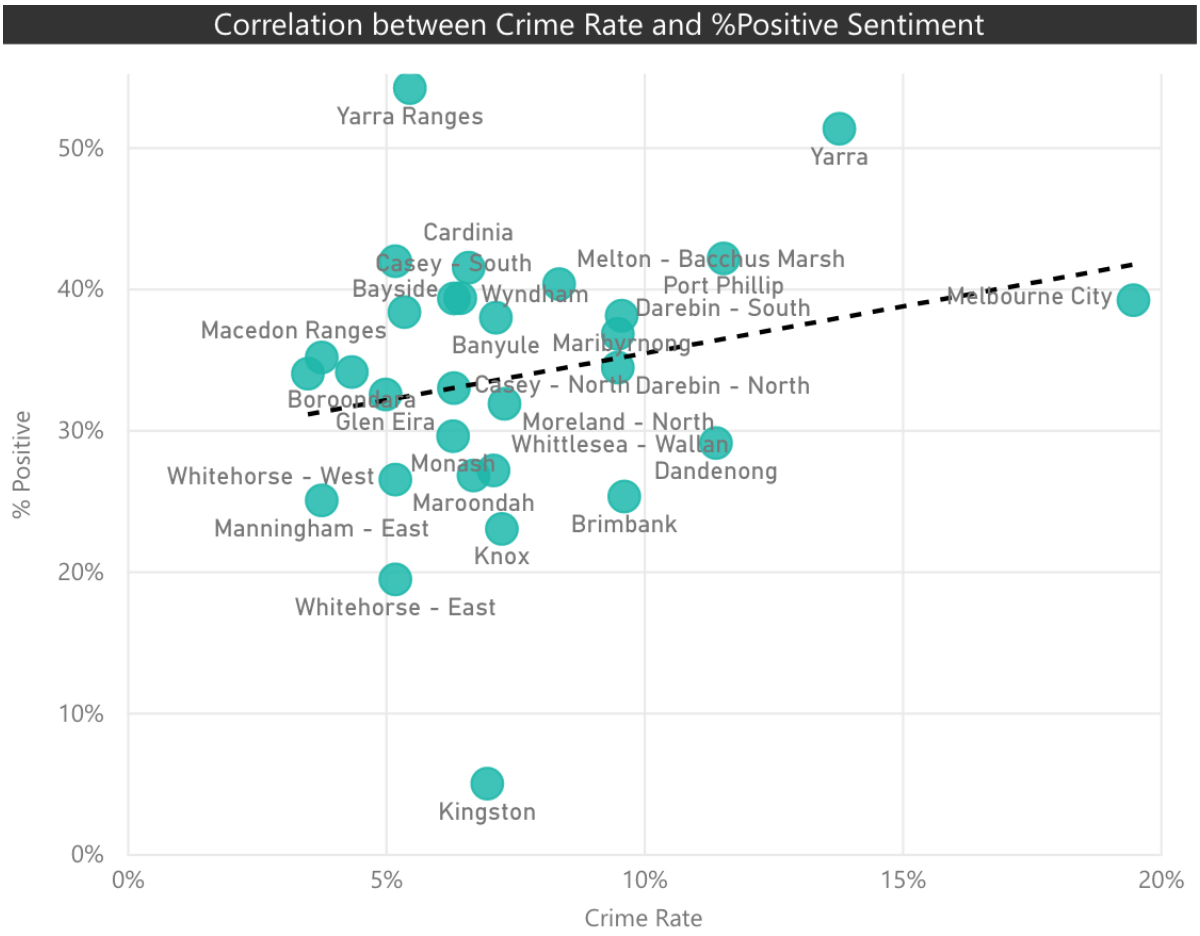

Figure 14: Pie chart and Bar chart of sentiment and party inclination vs sentiment

Scenario 4

Are people happier who live in areas that have the highest housing prices?

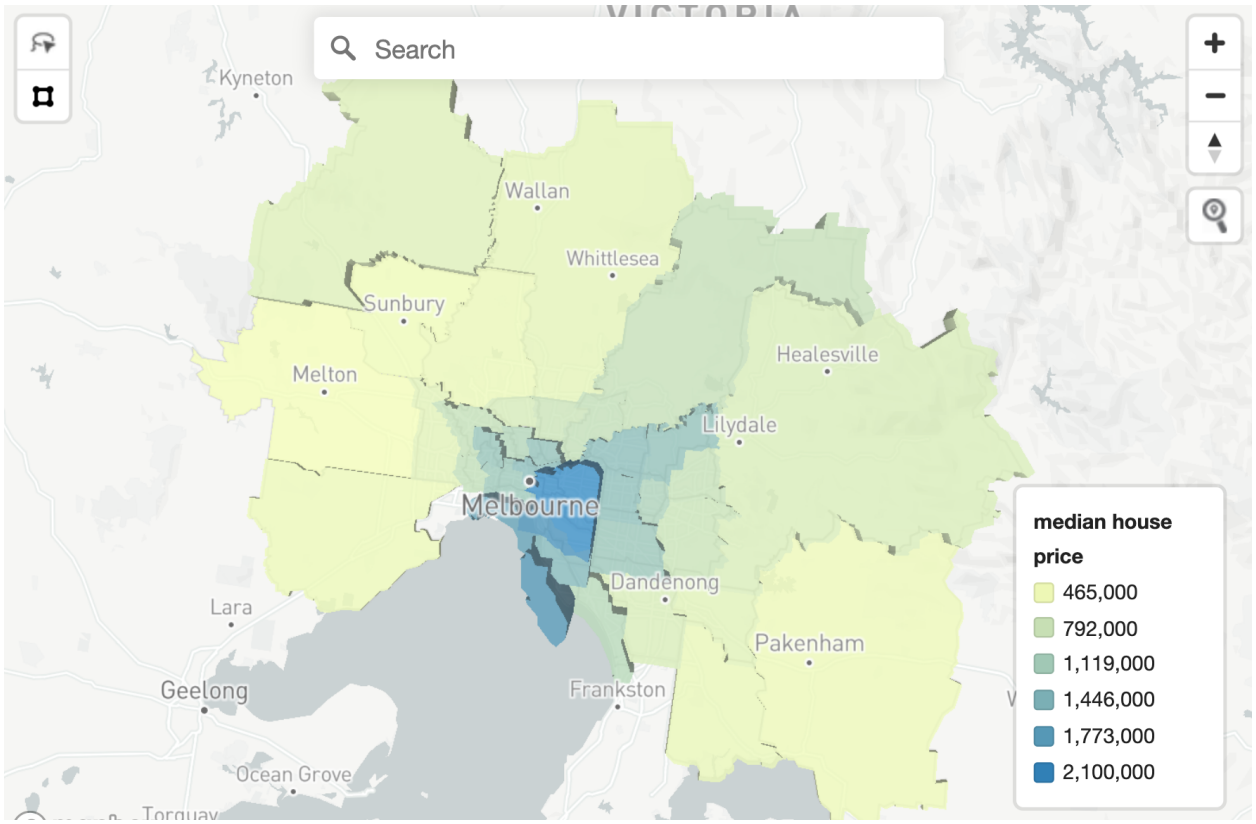

Figure 15: Height map of median house prices

The figure above (Figure 15) shows the median house price per SA3 region (the darker the colour, the higher the median house price). For this scenario, we used the general sentiment of each region to interrogate whether there is a correlation between housing prices and the population's happiness. From the scatter plot below, it can be seen that there is almost no correlation between the two variables. This could be attributed to the fact that people who live in regions with higher housing prices are well-established in their lives and are financially stable, resulting in positive sentiments in these regions. On the other hand, affluent people may be exposed to a variety of stresses (from their jobs, for example) which may result in negative sentiments. The interplay between the sources of positivity and negativity is potentially responsible for this weak correlation between housing prices and sentiment within a region.

Additionally, as mentioned in the previous scenario, the inherent demographic limitation imposed on tweets hinders our analysis of what a region's happiness level truly is.

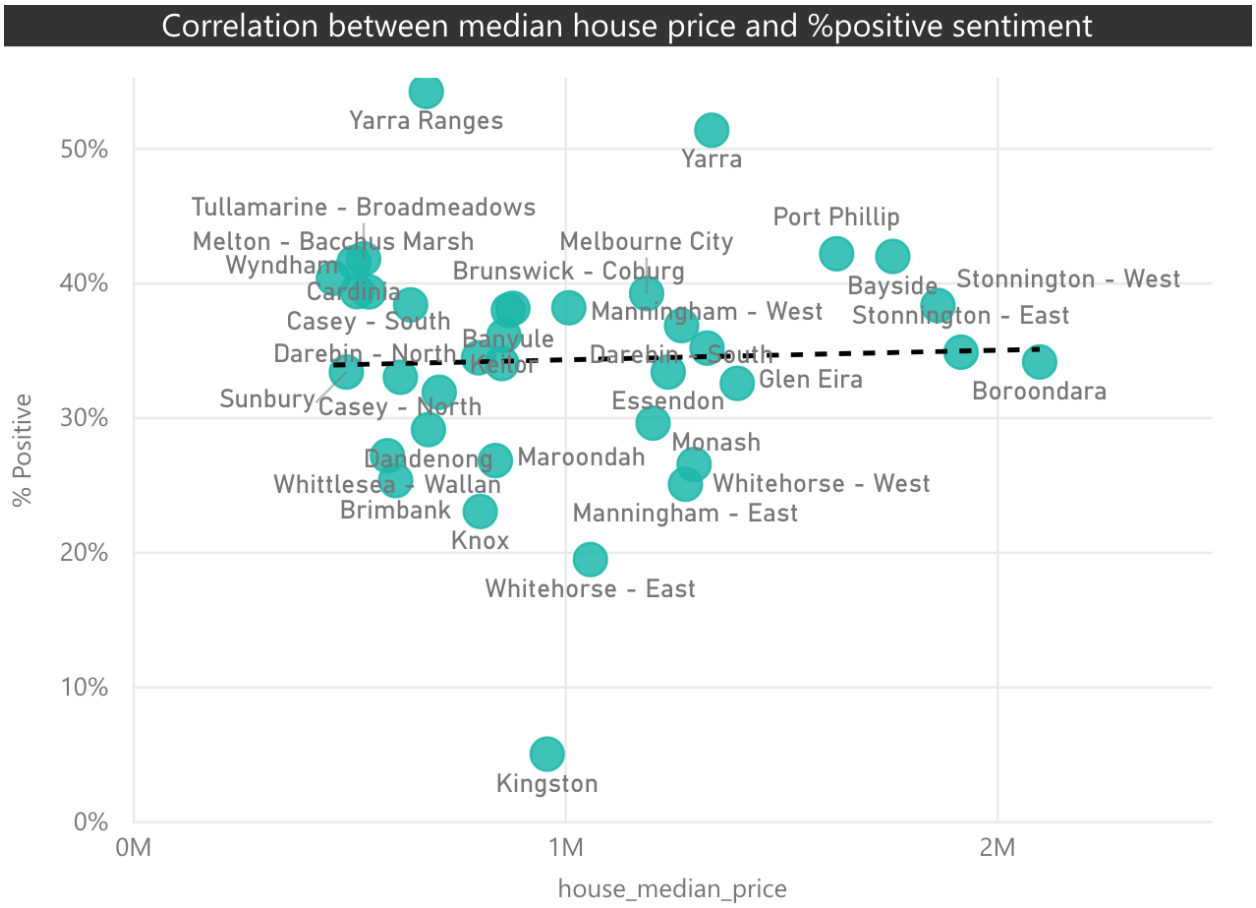

Figure 16: Plot of median house prices vs % positive tweets

SA3	# tweets collected	median house price	% Positive
Banyule	385	870,000	37.92%
Bayside	199	1,760,000	41.92%
Boroondara	224	2,100,000	34.08%
Brimbank	92	609,000	25.27%
Brunswick - Coburg	413	1,010,000	38.11%
Cardinia	165	512,750	41.46%
Casey - North	429	620,000	32.94%
Casey - South	29	547,333	39.29%
Dandenong	383	685,000	29.06%
Darebin - North	411	801,000	34.39%
Darebin - South	417	1,271,000	36.78%
Essendon	409	1,240,000	33.33%

Figure 17: List of median house prices, tweets collected and % positive tweets

Issues and Difficulties

Numerous difficulties were encountered while performing analysis on our scenarios. As you might have noticed in the Results section, the data used for the chosen scenarios are from different years and sources, and may introduce inconsistencies and inaccuracies into the results and correlations between various variables. Another difficulty, already mentioned in the Results section, is that evaluating the general sentiment of a region solely from tweets might be too reductive, as social media platforms do not tend to reflect the true demographic of a population.

Truncated tweets in the bigTwitter.json file were also problematic as this could have caused the sentiment evaluation of certain tweets to be inaccurate.

Occasionally, the datasets obtained from AURIN, would contain missing data and this would hinder analysis as we would have to either manually fill in the data point with data from another

row with the same or the most similar information, or delete the row with the missing data altogether.

Error Handling

Twitter

There are several errors that could be encountered when working with the Twitter API. It is important to deal with these errors as it could potentially lead to a ban of the user's twitter developer account.

Error codes such as 420 - *Enhance Your Calm* and 429 - *Too Many Requests* are quite common. Error code 420 is returned when an app is rate limited for making many requests and thus surpassing the permissible limit. Error code 429 is returned when the requests cannot be processed as the rate limit of the app has been exceeded. If these error codes are returned multiple times, the developer account could be locked.

In addition to this, we faced a `'Connection broken: IncompleteRead'` exception when running the Stream API for a long time. In order to deal with this error, we had to catch this exception as `ProtocolError` and continue the process.

Database

When dealing with large databases, one can expect to encounter duplicate documents at any time. To deal with removal of duplicate tweets, we avoid creating duplicate tweets in the first place. With CouchDB's flexibility of renaming document's id and its feature that enforces each document to have a unique document id, we rename the document id as the tweet id and when loading a tweet in the database, we check if a document already exists with the same document id as the tweet id. This is particularly useful when using the Stream API and the Search API simultaneously.

Helpful Links

GitHub repository: https://github.com/s3554374/CCC-AS2

Video demonstration of the project solution on YouTube:

- Ansible Deployment: https://youtu.be/rqgQlEKxzzc
- Harvester: https://youtu.be/xjzQnvtadxg
- Mapreduce views: https://youtu.be/lUMzoGAE184
- Front end website: https://youtu.be/vsbD5HOyDOQ

Website: 172.26.131.188 (if cluster-1 instance has been preserved as mentioned in the User Guide). Otherwise, this will depend on the cluster's IP address, which will be output at the end of the Ansible script.

References

AURIN Home. (2021, May 24). AURIN. Australian Urban Research Infrastructure Network. https://aurin.org.au/

Crime Statistics Agency Victoria. (2019, December 31). Https://Www.Crimestatistics.Vic.Gov.Au/. https://www.crimestatistics.vic.gov.au/

Twitter API Documentation. (n.d.). Docs | Twitter Developer Platform. Retrieved May 26, 2021, from https://developer.twitter.com/en/docs/twitter-api

Overview — Apache CouchDB® 3.1 Documentation. (n.d.). CouchDB Documentation. Retrieved May 26, 2021, from https://docs.couchdb.org/en/stable/

Docker Desktop. (n.d.). Https://Docs.Docker.Com/Desktop/. Retrieved May 26, 2021, from https://docs.docker.com/desktop/

Configure scheduled refresh - Power BI. (2021, April 16). Microsoft Docs. https://docs.microsoft.com/en-us/power-bi/connect-data/refresh-scheduled-refresh

On-premises data gateway - Power BI. (2019, July 15). Microsoft Docs. https://docs.microsoft.com/en-us/power-bi/connect-data/service-gateway-onprem

Robinson, L. (2021, May 13). *Social Media and Mental Health*. HelpGuide.Org. https://www.helpguide.org/articles/mental-health/social-media-and-mental-health.htm