Big Data analysis of green spaces' effect on public sentiment of COVID-19 in Melbourne during restrictions

Jakob Van Dyk

Student Number: s3923216

1. Introduction

1.1. Problem statement

The COVID-19 crisis highlighted the importance of open green spaces as restrictions tightened and social isolation became the norm. More than ever, green spaces were used for exercise, health, well-being, relaxation and respite from the confines of a person's home. This is especially true in the built-up suburbs of the metropolitan Melbourne area, where the size of properties is small and access to open spaces are limited. To understand how public sentiment was affected by access to green spaces during restrictions, we will use GIS (Geographic Information Systems) to spatially analyse Twitter data overlaid with green spaces. Social media platforms such as Twitter can serve as a lens to view collective experiences, with the benefit of being a readily available raw dataset containing geographic location, time and textual information. This is a more efficient means of research than alternatives, like canvassing park visitors. The results of this research will help inform decision-makers of better implementation of urban planning.

1.2. Literature review

Green space has proven positive impacts on improving mood. Self-reported measures of anger and sadness were found to be improved after exposure to the natural environment compared to the built environment (Bowler et al., 2010). Distance from green space was also found to be associated with lesser counts of anxiety and mood disorder treatments (Nutsford et al., 2013).

Social media platforms have been used to inform research with unique socially generated information about sentiment in situational contexts. The benefit of this method is sidestepping sample size and cost limitations of traditional means of collecting data, such as surveys. Research into the correlation between social media expressions and happiness found that social media data could be utilised to estimate attitude patterns at the population level (Mitchell et al., 2013).

Pertinent to this proposal, tweet data have been used to analyse the effects of green space on mental health in Melbourne (Hu and Sinnott, 2019). While overall finding no significant relationship between sentiment and green space, suburbs with large amounts of green space were found to be correlated with sentiment. Due to the lack of outlets and activities during lockdowns, as well as increased social media use, the relationship between sentiment and green spaces has likely shifted since 2019. This specific situational context requires additional research. Furthermore, this study buffered circles around the centroids of green spaces at a

constant value of up to 400m. Considering the number of green spaces in Melbourne, and the variety of sizes they come in, the buffering circles could envelop all of Melbourne or not fully cover the extent of a large green space. The limitations of this study, therefore, compel a new methodology.

Similar research of twitter data to measure public attitudes has been carried out across the world. Analysis of Twitter data in London on parks and visitors was conducted to extract spatiotemporal patterns of park visitors and resulting textual sentiment (Kovacs-Györi et al., 2018). Concluding results indicated that tweets were sent in parks 3-4km from their centre of activity and were more positive than elsewhere. In Massachusetts, Twitter data revealed higher sentiment scores were found in tweets located in commercial and public areas on weekends and during noon/evening (Cao et al., 2018). Geo-tagged tweets about COVID-19 were used to identify trends of sentiment disparity across the top 10 most populated North-American cities (Melotte and Kejriwal, 2021). Tweets over 10 months in 120 areas of Kuwait were analysed for primary topics of interest, the spatiotemporal distribution of these interests and spatial patterns of these distributions (G. Almatar et al., 2020). The study found emotional expressions were most common on weekends while the religious and political discussion was prominent on weekdays and spatial clustering of topics occurred across days of the week.

1.3. Aim and objectives

Aim:

This research aims to create a new dataset of web-scraped spatiotemporal data from Twitter users in Melbourne, to analyse and evaluate the effect of green spaces on public sentiment about COVID during restrictions. The results of this analysis will allow the government to better plan and improve public space and infrastructure to improve quality of life. Commercial organisations can also utilise the information to maximise utility in business operations such as location and opening hours. To achieve the aim, the objectives are set up as follows:

1. To derive a spatiotemporal dataset from tweets with quantified expressed sentiment about COVID-19 in Melbourne during lockdown restrictions;
2. To identify tweets with high and low scoring sentiments;
3. To spatially analyse with an overlay of green spaces;
4. To identify if proximity with green spaces correlates with higher sentiment scores;
5. To provide implications for decision makers and stakeholders to better utilise green space during restrictions and post-pandemic.

2. Data Models

This research will use a vector data model for analysis. Twitter records geo-location data at a single point with coordinates. These points represent the relative coordinate pairs of where a tweet was sent. Tweets are discrete objects and are not space filling. As such a vector model is the best suitable data model to represent their location. The Greenspace Dataset similarly utilises a vector model, using polygons to represent Parks, Reserves, National Parks, Conservation Areas, Forest Reserves, Recreational Areas and Open Space.

3. Data Collection

3.1. Study area

The study area will be limited in scope to metropolitan Melbourne. This is defined by the Victorian government as a collection of 31 local councils (VICGov, 2021). The purpose of restricting the study to this area is due to the Victorian government's dichotomy of regional Victoria and metropolitan Melbourne. Metropolitan Melbourne has been under lockdown for over 200 days as of writing, while regional Victoria has had restrictions to a relatively lesser extent.

To define this within GIS, a shapefile of a polygon showing the outer boundary of Plan Melbourne's Metropolitan Region will be imported. The shapefile was soured from the Department of Environment, Land Water and Planning, contains the 31 local councils making up metropolitan Melbourne (DELWP, 2017). Figure 1 below demonstrates the extent of the study area.

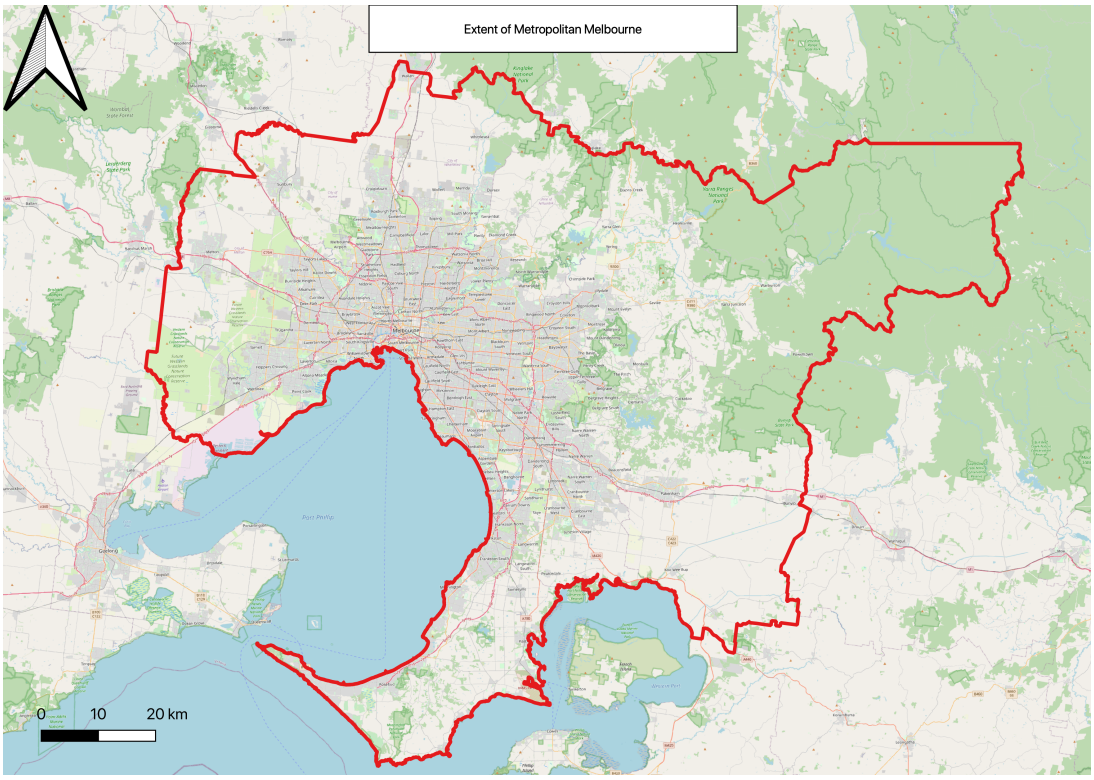

Figure 1. *Extent of Metropolitan Melbourne*

3.2. Datasets and sources

The dataset containing geometry and metadata of green spaces is sourced from the Australian Urban Research Infrastructure Network, provided by the Public Sector Mapping Agency, now known as Geospatial Australia (PSMA, 2020). The green spaces of an area are represented by polygons and is downloadable as a shapefile. Figure 2 below gives a visualisation of the Greenspace dataset around Carlton.

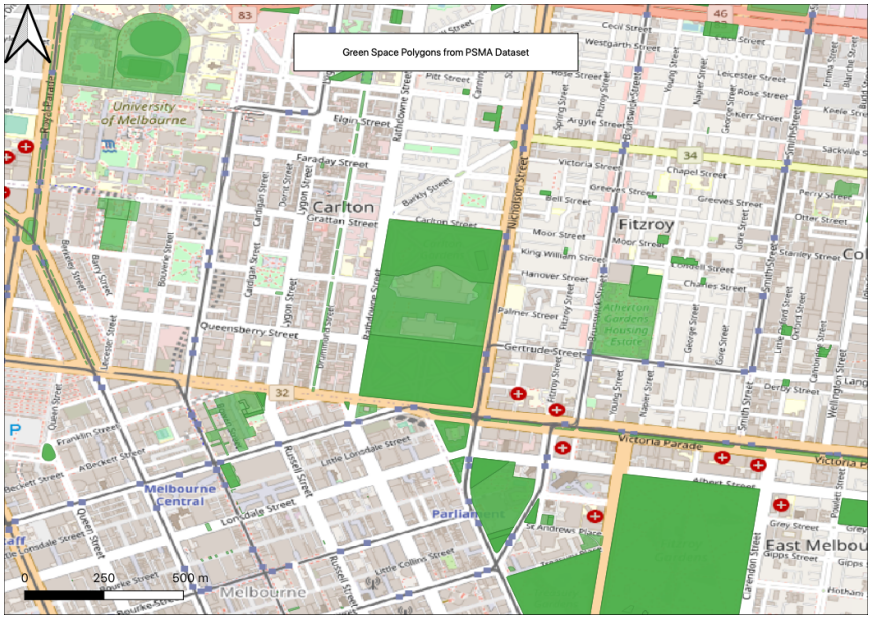

Figure 2. *Visualisation of Greenspace dataset around Carlton*

For developing a geographically restricted dataset of Twitter information, an open-access daily updated super-dataset is used. Since the 20th of March 2020, the *GeoCOV19Tweets Dataset* has collected geo-tagged Tweet IDs in English, which reference COVID-19 from around the world (Lamsal, 2020).

Twitter's content distribution policy does not allow for researchers to share the majority of information about user's tweets, so the dataset contains only the tweet ID. Information retrievable from Twitter relevant to the study includes time, location, tweet ID and text contents.

The dataset does contain processed sentiment analysis, however, which was achieved through the TextBLOB library. The tweets are pre-processed to clean the text of the hash symbol (#), the user mention symbol (@), URLs, extra whitespace and paragraph breaks (Lamsal, 2021).

TextBLOB then was used to score sentiment polarity as a continuous value from -1 to +1. This is advantageous as it defines sentiment on a relative basis rather than as discrete qualitative categories (strongly positive, slightly positive, neutral etc.). Positive sentiment ranges between 0 and +1 while negative sentiment ranges between -1 and 0, with 0 representing neutral sentiment.

3.3. Data management

Since the dataset only includes tweet ID's, to access the metadata from tweets they first need to be "hydrated". To do this a request is made from the Twitter v2.0 API using the tweet ID as the input. Twarc 2, a python library managing twitter requests, was used to access information from the Twitter API and hydrate the Tweet IDs (SamHames, 2021). The geo-tagged COVID-19 dataset contained processed sentiment scores, which needed to be stripped to use Twarc. This was done through Google sheets, as Microsoft Excel truncates numbers beyond the 13 digits of the IDs.

Filtering of the tweets then needs to be conducted to ensure only data within the scope of the research is present. Only tweets with known and exact coordinates are included in the dataset while those with null values are discarded. Tweets that make it through this filtering process are then output as a JSON file, the format that Twitter uses for its data storage. To make it more human-readable and accessible in GIS, the JSON file is converted to CSV. The file can now be imported into GIS software and the tweets plotted as points using the field "coordinates 0" for longitude and "coordinates 1" for latitude.

Next, the points are filtered based on the coordinates to ensure they are within the polygon of the study area described in Figure 1. This could be done earlier through python coding, but here will be computed using software to demonstrate the spatial

analysis capabilities of GIS. Figure 3 below summarises the data flow of tweets for this research.

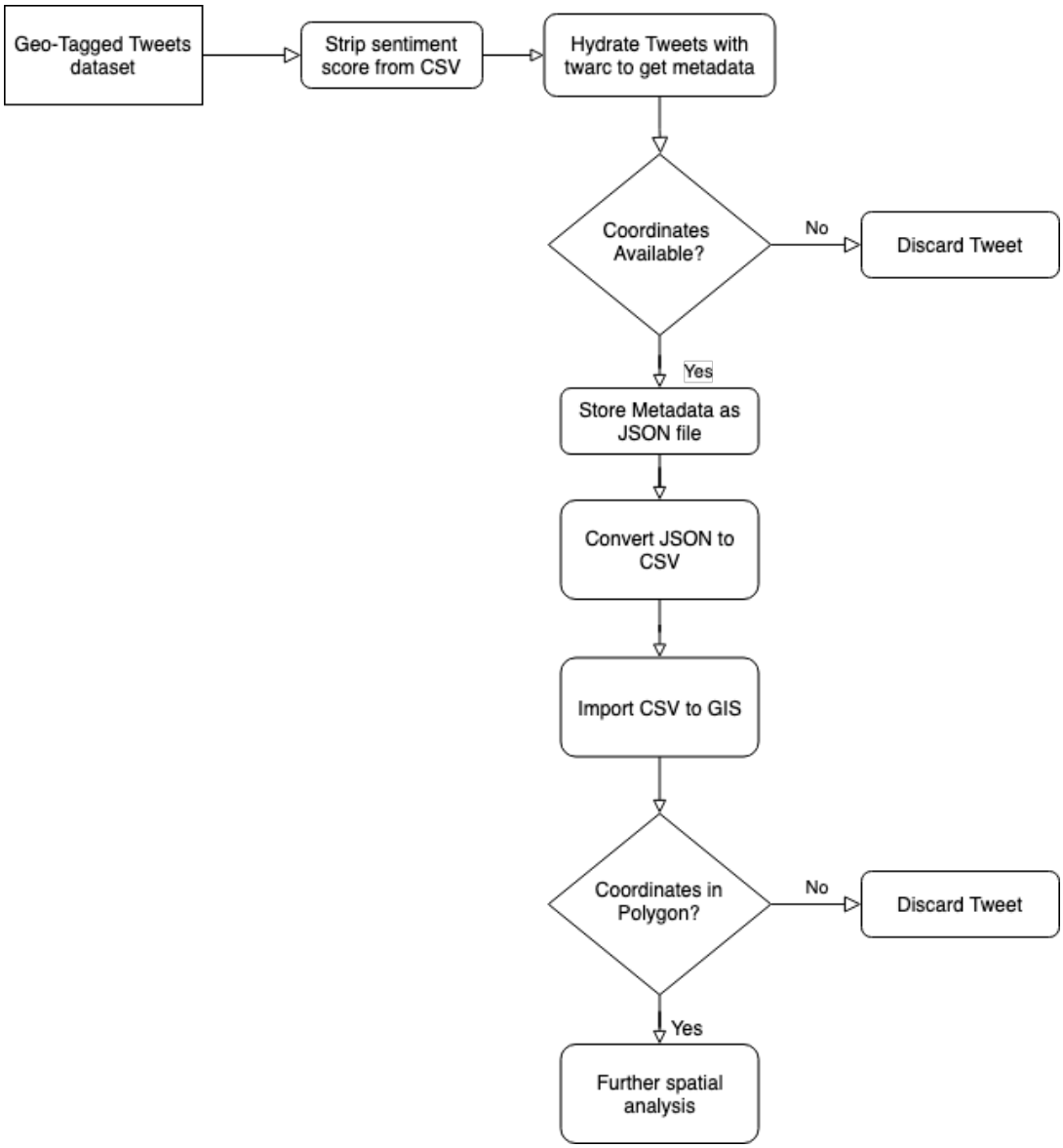

Figure 3. *Data flow of tweets*

4. Data Analysis and Expected Results

4.1. Spatial analysis

OpenStreetMap will be used for the visualisation of city features. The Greenspace dataset will be added to the project and a bounding box created around it. Points of tweets will be created from coordinates in the attribute table. All layers will be reprojected into the project CRS of EPSG: 28355 - GDA 94/MGA Zone 55 to ensure congruency, the accuracy of results, and creating buffers in metres. To spatially analyse, first, a semi-line sweeping algorithm will be used to filter for points within the polygon of the study area specified in Figure 1. A topological overlay of Twitter coordinates with Greenspace polygons will be used to identify any relationships or patterns between the datasets.

The plotted points of tweet coordinates will be buffered with three rings at a constant distance of 50m, in part to represent the uncertainty of smartphone location. Furthermore, the variable distance ring buffers allow for the discernment of any change in sentiment due to distance from a green space. Intersect analysis will be used to identify if a buffered ring's area intersects with a green space polygon. When multiple ring areas intersect with a polygon or multiple polygons, the closest ring with the highest percentage of area identified in the table will be selected.

If there is a pattern of sentiment being higher in the inner circles of the buffered tweet than outer circles and beyond, it can be extrapolated that green spaces have

a positive effect on the sentiment expressed about COVID-19. Depending on the distance, the sentiment score will be weighted. By comparing the sentiment score of tweets of a buffer that intersects with a green space polygon against points tweets that do not, we can spatially analyse the difference that green space makes on expressed attitudes of COVID-19.

4.2. Maps and tables

A visualisation of the spatial analysis is given below as a proof of concept in Figure 4. The tweet is located in Richmond at the Dame Nellie Melba Memorial Park, was created on the 31st of August 2021 at 09:18:43 pm UTC and had a sentiment score of 0, representing neutrality.

Figure 4. *Spatial analysis proof of concept*

Examples of intersect analysis, used to determine which buffer ring area is used for further analysis, is given below:

Figure 6. *Example of table used to determine appropriate distance of tweet from green space*

ringId	distance	psma_greenspace_polygon_202008-_area	psma_greenspace_polygon_202008-_pc
3	150.000000	2490.467458882733	6.342991312182383
1	50.000000	2067.85976617477	26.334545673588426
3	150.000000	1331.3586804652587	3.390932999131022
2	100.000000	340.87784131010994	1.446974123597342
2	100.000000	211.00898114882875	0.8957707824113448
2	100.000000	211.00898114882875	0.8957707824113448
3	150.000000	195.5807838526962	0.4981651984993275
3	150.000000	195.5807838526962	0.4981651984993275
2	100.000000	145.89208572510688	0.6193201874648181

Further statistical analysis to determine the significance and strength of effect of green space on sentiment, along with other factors, will be conducted through multi-linear regression. A full description of this process is outside the purpose of this research proposal.

5. Discussions

5.1. Related issues

Tweet sentiment about COVID-19 could be negatively affected due to new and unexpected announcements of COVID-19 outbreaks and further restrictions. While a Twitter user may be more positive about COVID-19 by being in a green space relative to others, the sentiment could still be lower than the overall average of

sentiments. Due to the timestamp that is associated with every tweet, however, we can account for any outliers due to specific events and restrictions.

Other variables that are statistically significant in measuring expressed sentiment about COVID-19 include; socioeconomic status, poverty rate, homelessness, housing stress, loss of employment, population density and percentage of English tweets of all tweets (Hu and Sinnott, 2019). These variables' effects on sentiment will be compared with the distance from green space to relatively position the factors based on strength.

5.2. Limitations

One limitation is the original dataset of geo-tagged COVID-19 tweets. Twitter only lets developers to stream 1% of tweets as they are posted, so they are a sample of all tweets about COVID-19. The amount of geo-tagged tweets is also relatively small, with 421,210 out of 1,603,107,609 tweets collected having coordinate information.

There are also gaps in the dataset, where data starts from the 20th of March 2020, and the 29th to 30th of March are missing due to technical issues. All other months have data for every day. Sentiment analysis is similarly missing from the 27th to the 28th of October 2020. These gaps will not be considered in the dataset.

Another limitation is the demographic diversity of Twitter users. Twitter users are likely to be younger, earn, and have higher education (Sehl, 2020). Similarly, only tweets in English were included in the data set, so only inferences about English speakers can be made. This limits the generalisations able to be made from the results.

The accuracy and precision of recorded coordinates by smartphones also limit the results. Availability of satellites, the multi-path error caused by the reflection of signals off structures in urban canyons and the quality of the receiver can all cause deviation of recorded coordinates from their true location.

5.3. Impact and implications

The main benefit of this study will be a quantifiable evaluation of the value green space provides for the well-being of the public. By understanding how green space affected Melbourne's general public's sentiment towards COVID-19, decision-makers can make informed choices about land usage and land planning for the benefit of society.

As a secondary benefit, commercial businesses could find the time and place where customers had the highest sentiment, even in times of duress, to maximise profitability through property choice and opening hours.

6. Conclusion

As society moves away from uninhibited industrialisation and seeks to find equilibrium with the natural environment, quantified research is needed to ensure the right balance is struck. This is necessary for decision-makers to improve the liveability of cities and maximise utility, through appropriate land planning and usage. This proposed research uses big data sourced from Twitter, to measure the effect green spaces had on sentiment during COVID-19 restrictions, a time of great mental stress. From the results of this research, decision-makers will be able to ensure the balance of urban and natural is met correctly and improve resident's quality of life. Secondary to this, businesses may be able to use the dataset and results to inform their operations. Through a big data analysis of tweet sentiment during COVID-19, we can measure the value that green spaces add to Melbourne's liveability.

Bibliography:

BOWLER, D. E., BUYUNG-ALI, L. M., KNIGHT, T. M. & PULLIN, A. S. 2010. A systematic review of evidence for the added benefits to health of exposure to natural environments. *BMC Public Health,* 10**,** 456.

CAO, X., MACNAUGHTON, P., DENG, Z., YIN, J., ZHANG, X. & ALLEN, J. G. 2018. Using Twitter to Better Understand the Spatiotemporal Patterns of Public Sentiment: A Case Study in Massachusetts, USA. *International Journal of Environmental Research and Public Health,* 15.

DELWP 2017. Metropolitan Region. *Plan Melbourne Shape Files.*

G. ALMATAR, M., ALAZMI, H. S., LI, L. & FOX, E. A. 2020. Applying GIS and Text Mining Methods to Twitter Data to Explore the Spatiotemporal Patterns of Topics of Interest in Kuwait. *ISPRS International Journal of Geo-Information,* 9.

HU, Y. & SINNOTT, R. O. Big Data Analytics Exploration of Green Space and Mental Health in Melbourne. 2019 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), 14-17 May 2019 2019. 648-657.

KOVACS-GYÖRI, A., RISTEA, A., KOLCSAR, R., RESCH, B., CRIVELLARI, A. & BLASCHKE, T. 2018. Beyond Spatial Proximity—Classifying Parks and Their Visitors in London Based on Spatiotemporal and Sentiment Analysis of Twitter Data. *ISPRS International Journal of Geo-Information,* 7.

LAMSAL, R. 2020. Coronavirus (COVID-19) Geo-tagged Tweets Dataset. IEEE Dataport.

LAMSAL, R. 2021. Design and analysis of a large-scale COVID-19 tweets dataset. *Applied Intelligence,* 51**,** 2790-2804.

MELOTTE, S. & KEJRIWAL, M. 2021. A Geo-Tagged COVID-19 Twitter Dataset for 10 North American Metropolitan Areas over a 255-Day Period. *Data,* 6.

MITCHELL, L., FRANK, M. R., HARRIS, K. D., DODDS, P. S. & DANFORTH, C. M. 2013. The geography of happiness: connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one,* 8**,** e64417-e64417.

NUTSFORD, D., PEARSON, A. L. & KINGHAM, S. 2013. An ecological study investigating the association between access to urban green space and mental health. *Public Health,* 127**,** 1005-1011.

PSMA 2020. PSMA - Transport & Topography - Greenspace (Polygon) August 2020. *In:* AUSTRALIA, P. A. L. T. A. G. (ed.). https://portal.aurin.org.au.

SAMHAMES 2021. Twarc. Documenting the Now.

SEHL, K. 2020. *Top Twitter Demographics That Matter to Social Media Marketers* [Online]. Hootsuite. Available: https://blog.hootsuite.com/twitter-demographics/ [Accessed 2021].

VICGOV. 2021. *Metropolitan Melbourne* [Online]. Available: https://liveinmelbourne.vic.gov.au/discover/melbourne-victoria/metropolitan-melbourne [Accessed 2021].