An Initial Visual Analysis of the Relationship between COVID-19 and Local Community Features

1st Jie Hua
Faculty of Information Engineering
Shaoyang University
Shaoyang, China
Jie.Hua@alumni.uts.edu.au

2nd Mao Lin Huang
School of Computer Science
University of Technology Sydney
Sydney, Australia
Mao.Huang@uts.edu.au

3rd Chenglin Zhao
Faculty of Information Engineering
Shaoyang University
Shaoyang, China
Zhao.C.L@outlook.com

4th Shuyang Hua
Faculty of Engineering
University of Sydney
Sydney, Australia
Shua6688@uni.sydney.edu.au

5th Catherine Shih
Faculty of Science
University of Sydney
Sydney, Australia
Cshi4928@uni.sydney.edu.au

2020 24th International Conference Information Visualisation (IV) | 978-1-7281-9134-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/IV51561.2020.00125

***Abstract*—Virus outbreaks are threats to humanity, and coronaviruses are the latest of many epidemics in the last few decades. In this work, we conduct a non-medical/clinical approach, generating graphs from five features concluded from the COVID-19 outbreak data and local community data in NSW (New South Wales), Australia, and offering insights from a visual analysis perspective. The results show that household income, population density and ethnicity affect the infection in different areas. Features such as human behaviours need to be imported for further COVID-19 research in the data science sector. This work is an initial step into this area and allows more insights to be brought into future COVID-19 research through a visual analysis perfective.**

Keywords—visual analysis; graph visualisation; graph drawing; coronavirus; COVID-19

I. Introduction

The recent COVID-19 outbreak has infected 216 countries, areas or territories in the world as of 13 Jul 2020, causing 12,685,374 confirmed cases and 565,000 deaths [1]. To tackle this worldwide health crisis, medical/clinical research is crucial, along with studies from various perspectives, such as virus data analysis, etc., which may also assist in discovering deeper insights.

Some latest works have been conducted not only on the coronavirus itself but also on several potential impact features. Population density has been imported to further analyse its connection with the virus infection [2-4]. Household income has also been taken into account to investigate the impact of the virus outbreak by utilising transaction-level household financial data [5] and examining measurements in countries with low and middle incomes [6] in the COVID-19 epidemic. Ethnic classification systems have also been used to explore genetic and other population differences [8]. The U.K. (United Kingdom) is the first country in the COVID-19 surge with an ethnically diverse population, possibly contributing to the research of the disease's effects within different ethnic groups [9]. The works mentioned above provide more facts that need to be considered in COVID-19 analysis, yet, most of those are medical/clinical related, or only concern particular features.

This study is not related to medical/clinical research; it is an initial visual approach in COVID-19's data analytics. Based on LGA (Local Government Area) 's features such as income, infected source, population and ethnicity, etc., this study visually analyses such features and their relationships to the COVID-19 outbreak. Unlike existing works, this study only offers from a visual analysis perspective and addresses and combines several different elements to discover unknown patterns in the virus epidemic.

During the COVID-19 outbreak in NSW, Australia in 2020, our hypotheses are finalised based on infection case number in each LGA and LGA's community profiles. The four hypotheses below only consider relevant data in NSW.

- H1: Most cases were imported from overseas.
- H2: Infection case is population density related.
- H3: Infection case is household income related.
- H4: Infection case is ethnicity related.

The rest of this article is organised into several parts. In Section 2, relevant data and its processing step details are given, as well as the graph visualisation tools involved in experiments. In Section 3, we offer visual results and discuss outcomes in Section 4. Finally, we conclude our work and discuss future research in Section 5.

II. Methods

A. Data Processing

We downloaded and collected two types of data: COVID-19 case data that contains the location and likely source of infection [11]; LGA General Community Profile that includes relevant features such as population, gender, ethnicity, religious belief etc. [12]. In the cleansed dataset, we added features shown in Table I, and data were considered between 23/Mar/2020 and 07/Jul/2020. Only cases with confirmed LGA have been taken into account.

Eventually, we concluded 101 records that present 101 LGAs' case infection details and community profiles in NSW, Australia. Each record contains attributes shown in Table I.

TABLE I. Data Attributes

Name	*Description*
notification_date	Case reported date
postcode	Location's postcode
Source (including five fields shown in the right)	source-overseas; source-local confirmed; source-local not identified; source-interstate; source-under investigation

DOI 10.1109/IV51561.2020.00125

Name	*Description*
lhd_2010_code	Local Health District code
lhd_2010_name	Local Health District name
infected	The infected number on the date
Total_infected	Existing infected case number
LGA_code	Local Govern Area code
LGA_Name	Local Govern Area name
Median Household Income	Median Household Income of the LGA
Ethnicity (including many fields shown in the right)	Australian; Aboriginal; Chinese; English; Irish etc.
Area	The land area of the LGA
Population	The population of the LGA

B. Graph Visualisation

Tableau is a common visualisation platform that offers rich features to create interactive visual outcomes, and we applied bar chart, stacked bar chart, area chart, line chart in Tableau [10, 13-16] to transform raw data into information, to hence explore and understand the complex data through visual observations.

C. Procedure

The proposed approach uses the raw data finalised from the data processing step as the input for Tableau; then generates graphs from features such as infection case number, ethnicity, population, income, and infection source etc.; the charts are then observed to find out if the hypotheses are supported.

In the experiments, four ethnicities were selected: Australian, English, Chinese and Irish, as they are the top four ethnicities in the greater Sydney area from 2016 Census statistics.

Throughout this manuscript, the year is 2020 if only day and month are mentioned; Sydney represents Sydney CBD; ethnicity refers to cultural factors; for example, English implies immigrants from English- speaking countries.

III. RESULTS

Fig 1, 2, 3, 4 are presented with five features: infected case number, four ethnicities rate, household income, population density and infection source. All elements except the case number are fetched from LGA profiles in 2016 Census.

In Fig 1, the left y-axis indicates each LGA's infection number, the showing numbers are divided by 300, so as to put all elements in the graph onto the same level for observing purposes; the right y-axis represents four ethnicities' percentage in each LGA; the x-axis shows each LGA with infection cases (here it only represents LGAs with case number larger than ten); the grey area illustrates the infection case numbers, and from left to right, it is ordered from large to small. It can be observed that: Waverley has the most infection case till 07/Jul, followed by Sydney CBD; in the top nine LGAs with the most cases (those take 39.7% of total infection cases), there are high proportions of Australian and English; Irish's distribution tends to be steady; LGAs with more Chinese do not have significant infection cases, for example, there are 24% Chinese in Ryde with 70 cases, and 27.8% in Georges River with 59 cases; when there are more Australian and English, there are less Chinese, but Irish remain the same.

In Fig 2, the x-axis represents LGA groups with similar infection cases, and groups are divided based on case numbers: 1-49; 50-99; 100-149; and more substantial than 150. For example, the most left group comes with 50-99 case numbers. The top-left y-axis shows each ethnicity's percentage, and the top-right y-axis indicates the infection case number. Hence, there are 1,013 cases in the group of LGAs with 50-99 instances, English residents constitute 27.83% in those LGAs.

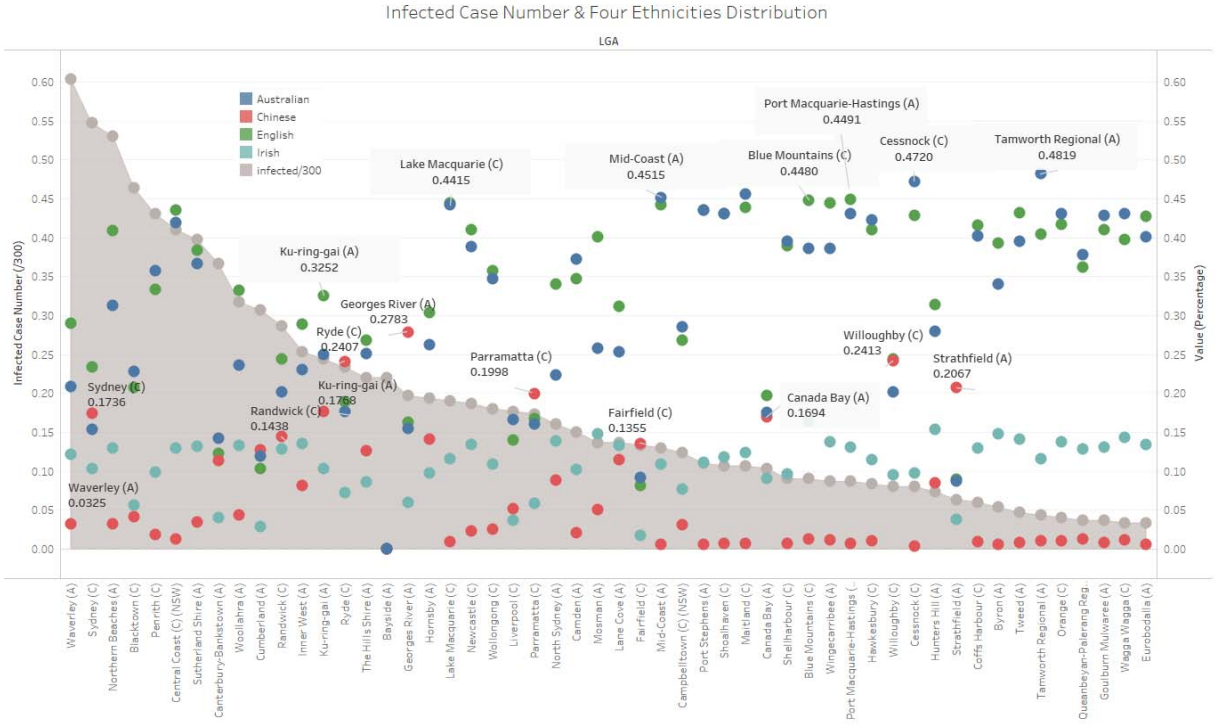

Fig. 1. LGA's existing infected case number and four ethnicities distribution (infection case number >= 10)

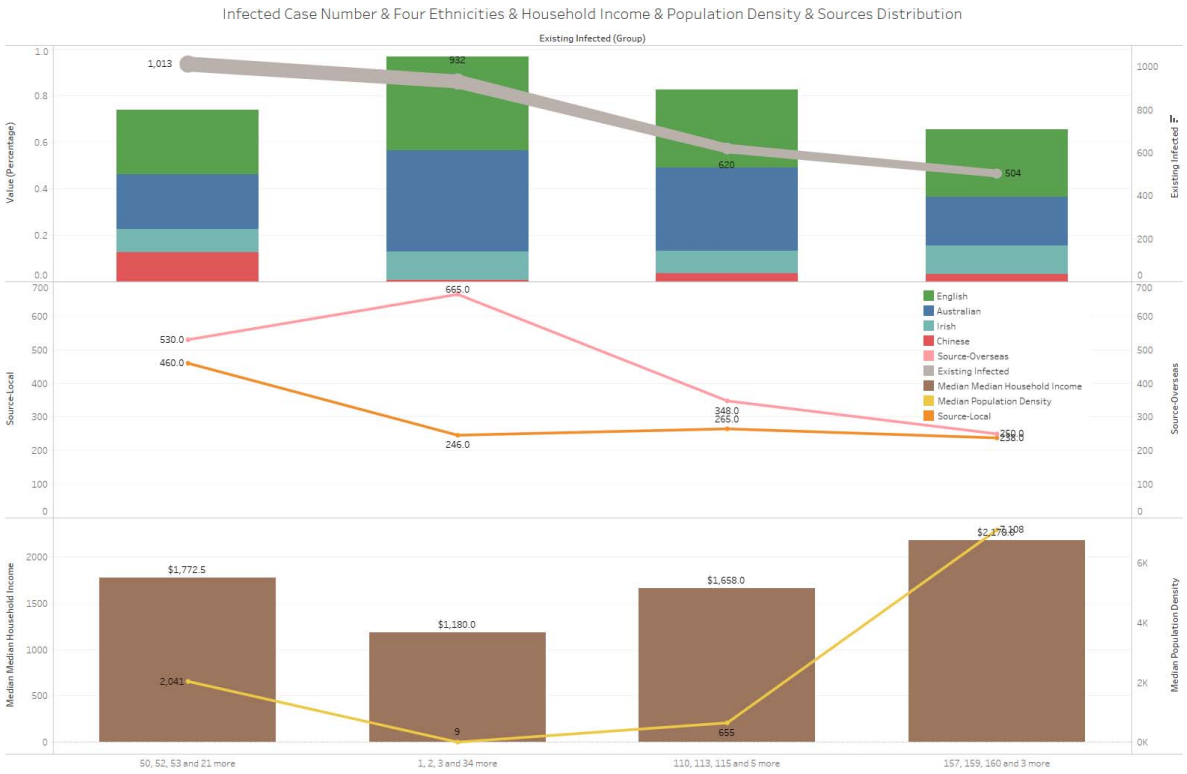

Fig. 2. Distribution of LGA's existing infected case (ethnicities, household income, infection source and population density) – four groups

The middle-left y-axis represents local-infected case number, and the middle-right y-axis characterises overseas-infected case number; the bottom-left y-axis signifies median household income, and the bottom-right y-axis describes median population density. The ratio of overseas import cases to local infection cases is 1.483:1; LGA group with 50-99 cases and 1-49 cases account for 33% and 30.4%, and the larger than 150 case group make up 16.4%; 50.6% cases occurred in LGAs with less than $1,658 household income, and 16.4% cases arose in LGAs with more than $2,178 household income (a roughly median household income in NSW); LGAs with higher ratio (2.7:1) of overseas import cases to local infection cases is the 1-49 group, which is also the LGAs with lower median household income. Detailed values of each feature are shown in Table II.

TABLE II. LGA GROUP FEATURES

Features	*Groups*			
	1-49	***50-99***	***100-149***	***>=150***
Total Infected	932	1,013	620	504
Local Infected	665	530	348	250
Overseas Infected	246	460	265	238
Australian (%)	43.27	23.32	35.72	20.85
English (%)	40.31	27.83	33.28	28.97
Chinese (%)	0.71	12.64	3.47	3.25
Irish (%)	12.33	10.03	9.83	12.17
Population Density	2041	2,041	665	7,108
Household Income	1,180	1,772.5	1,658	2,178

Fig 3 is similar to Fig 2, and only LGA groups are allocated with smaller scopes: 1-19; 20-39; 40-59; 60-79; 80-99; 100-149; >=150. We can see that: overseas import cases are more than local infection cases in most LGAs, the ratio of the infection sources in the high household income LGAs tend to be 1:1, which is similar as the situation in Fig 2; middle household income ($1,658-$1,916) LGAs form 48.8% cases, high household income (>$2,178) LGAs comprise 27.9% cases, and it is 23.4% in low household income LGAs. The LGA group with high population densities do not come with the most infection cases.

In Fig 4, the x-axis represents LGA with infection cases of more than ten; the left y-axis indicates each LGA's infection number; the right y-axis shows population density; the grey line illustrates total infection case numbers, the pink line demonstrates overseas infection case numbers, and the orange line displays the local infection case numbers. From left to right, it is ordered by the population density of each LGA. Some interesting facts are: the top two LGAs with the highest population density have the most infection cases, such as Waverley and Sydney, but others are not; LGAs with more cases do not always have higher population density, such as Blacktown, Northern Beaches, Penrith and Central Coast etc.; in most cases, LGAs have more overseas infection cases, except Ryde, Penrith etc.

IV. DISCUSSION

From all the results in Section III, we observed that: most infection cases were imported from overseas; the LGAs with the highest case numbers have higher population densities such as Sydney and Waverley, but other LGAs are not; the LGAs with lower population densities still have a chance to get more infection cases especially beaches with many tourists such as Northern Beaches; LGAs with more overseas cases

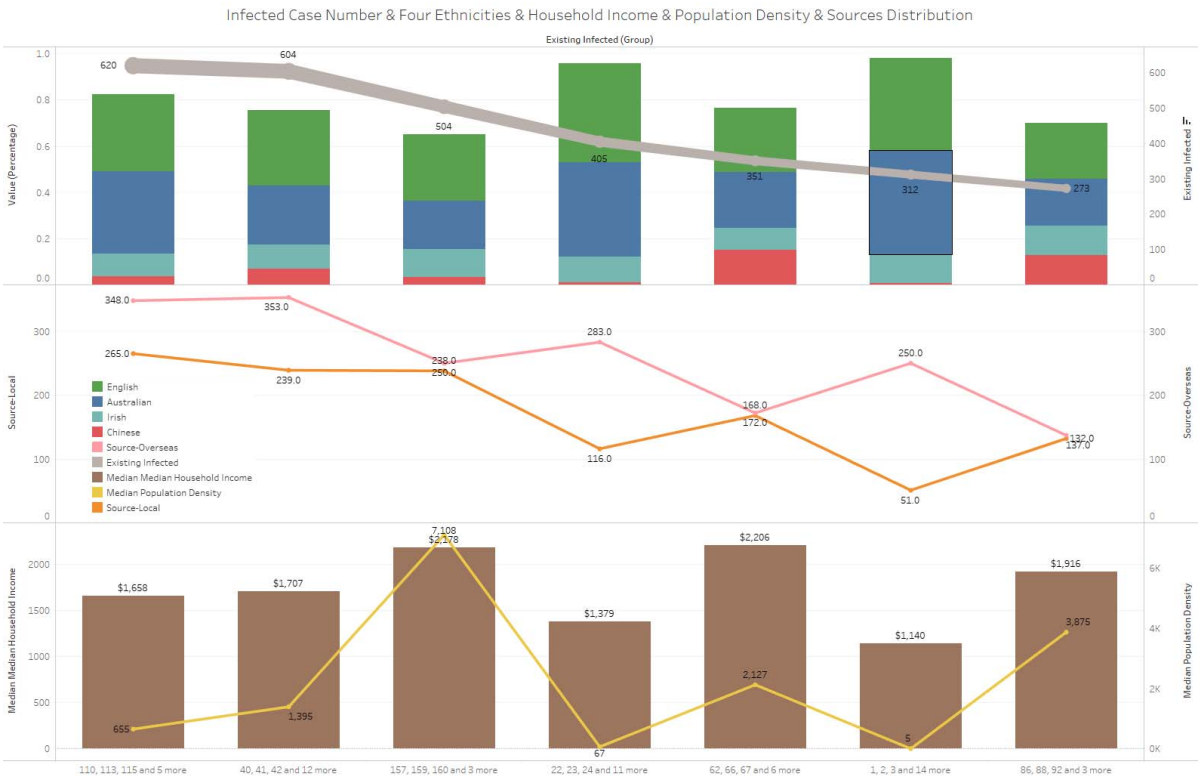

Fig. 3. Distribution of LGA's existing infected case (ethnicities, household income, infection source and population density) – seven groups

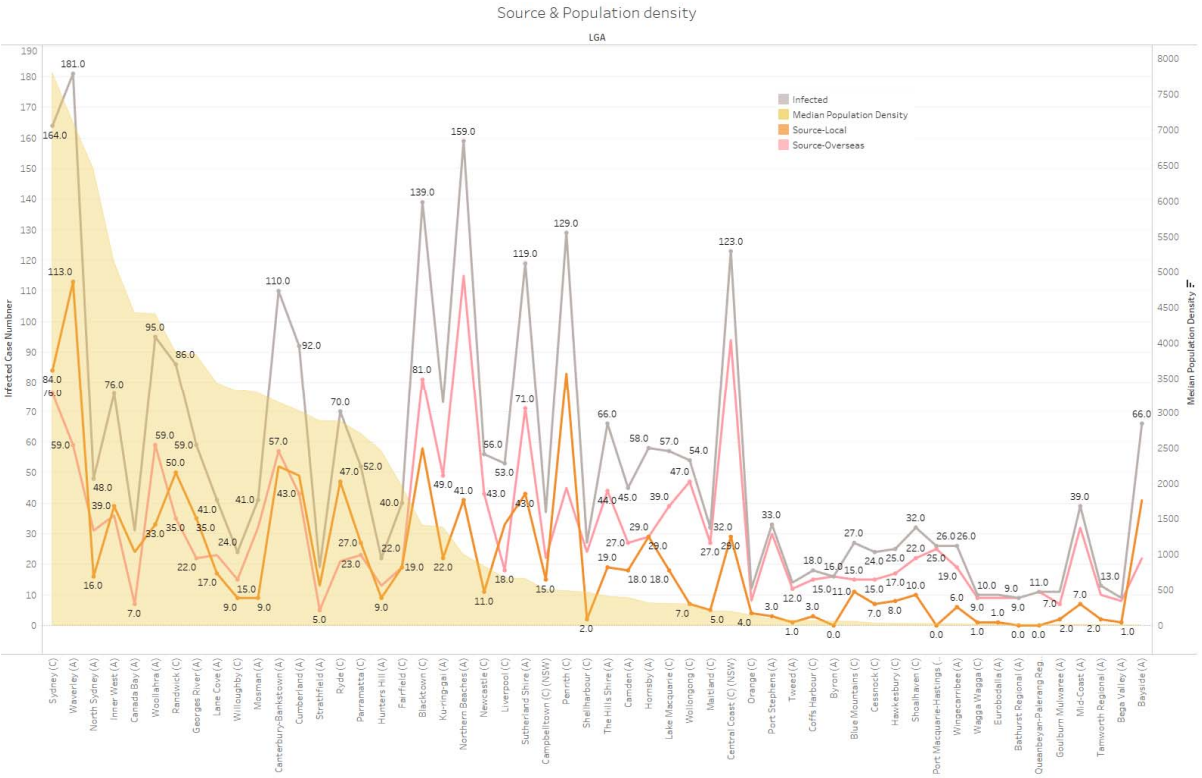

Fig. 4. Distribution of LGA's infected case (infection source and population density)

are not the wealthiest areas, most cases (63.4%) happened in LGAs with median household income; in those highly infected areas, LGAs with more Chinese tend to have fewer infection cases.

Concerning our hypotheses, we can conclude:

- For H1, in 88.1% LGAs, there are more overseas infected cases, and the ratio of overseas import cases to local infection cases in NSW is 1.483:1.
- For H2, population density has a connection to the infection case; however, there might be other factors that affect infection such as tourism [20] etc. We have not considered those in this study; hence we cannot prove the potential causes at this stage.
- For H3, LGAs with median household income are likely to have more infection cases, and most cases from overseas were transpired in LGAs with middle and low household incomes, the ratios of infection sources are inclined to be 1:1 in those LGAs with high household incomes.
- For H4, in NSW, the distribution of infection case in each LGA shows a potential connection to ethnicities. It might be caused by the difference of following the prevention measures such as less travel, avoiding crowded places, sanitising, and wearing a mask, etc. [17-19]. We have not done any research in those yet; thus, we cannot prove those possible reasons at this stage.

Although this is an initial step of our relevant research, some limitations remained need to be clarified. First of all, LGA profiles were collected from 2016 Census at Australian Bureau of Statistics, which may cause the accuracy issues of the analysing results; and there are many features have not been imported such as gender, dwelling structure, relationship in a household, method of travel to work, those all potentially affect the virus spreading; the current outcomes are mainly based on visual observing, lack of statistical analytics at this stage.

V. Conclusions

Through the experiments, we generated graphs for visual analysis of the COVID-19 outbreak from five major features. This work is not medical nor clinical related; all outcomes are established entirely on visual data analysis. Hence, this work is for people who have an interest towards statistical data rather than virology knowledge. Through this work, we can obtain some insights from the complex raw data via visual analysis, offering different views on COVID-19's research. In our future work, we could address the limitations mentioned and apply more features such as human behaviours, event timeline etc. to COVID-19 data analysis, discover unknown patterns in the COVID-19 outbreak from a visual analysis perspective.

References

[1] WHO. Coronavirus Disease 2019. 2020. Accessed on: 13 Jul, 2020. [Online] Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019

[2] J. Wang, K. Tang, K. Feng, and W. Lv, "High temperature and high humidity reduce the transmission of COVID-19," Available at SSRN 3551767. 2020.

[3] M. Ahmadi, A. Sharifi, S. Dorosti, S.J. Ghoushchi, and N. Ghanbari, "Investigation of effective climatology parameters on COVID-19 outbreak in Iran," Science of the Total Environment, p.138705, 2020.

[4] S. Feng, C. Shen, N. Xia, W. Song, M. Fan, and B.J. Cowling, "Rational use of face masks in the COVID-19 pandemic," The Lancet Respiratory Medicine, 8(5), pp.434-436. 2020.

[5] S.R. Baker, R.A. Farrokhnia, S. Meyer, M. Pagel, and C. Yannelis, "How does household spending respond to an epidemic? consumption during the 2020 covid-19 pandemic," (No. w26949). National Bureau of Economic Research. 2020.

[6] J. Hopman, B. Allegranzi, and S. Mehtar, "Managing COVID-19 in low-and middle-income countries," Jama, 323(16), pp.1549-1550. 2020.

[7] P. Lloyd-Sherlock, S. Ebrahim, L. Geffen, and M. McKee, "Bearing the brunt of covid-19: older people in low and middle income countries," 2020.

[8] M. Pareek, M.N. Bangash, N. Pareek, D. Pan, S. Sze, S, J.S. Minhas, W. Hanif, and K. Khunti, "Ethnicity and COVID-19: an urgent public health," 2020.

[9] K. Khunti, A.K. Singh, M. Pareek, and W. Hanif, "Is ethnicity linked to incidence or outcomes of covid-19?" 2020.

[10] J. Hua, G. Wang, M. Huang, S. Hua, and S. Yang, "A Visual Approach for the SARS (Severe Acute Respiratory Syndrome) Outbreak Data Analysis," International Journal of Environmental Research and Public Health, 17(11), p.3973. 2020.

[11] Data.NSW, NSW COVID-19 cases by location and likely source of infection. Accessed on: 13 Jul, 2020. [Online] Available: https://data.nsw.gov.au/data/dataset/nsw-covid-19-cases-by-location-and-likely-source-of-infection

[12] Australian Bureau of Statistics, Census. Accessed on: 13 Jul, 2020. [Online] Available: https://www.abs.gov.au/websitedbs/D3310114.nsf/Home/2016%20search%20by%20geography

[13] J. Hoelscher, and A. Mortimer, "Using Tableau to visualise data and drive decision-making," Journal of Accounting Education, 44, pp.49-59. 2018.

[14] I. Ko, and H. Chang, "Interactive visualisation of healthcare data using tableau," Healthcare informatics research, 23(4), pp.349-354. 2017.

[15] S. Hamersky, "Tableau desktop," Math. Comput. Educ. 50, 148. 2016.

[16] I. Datig, and P. Whiting, "Telling your library story: Tableau public for data visualisation," Libr. Hi Tech News, 35, 6–8. 2018.

[17] J. Xu, S. Zhao, T. Teng, A.E. Abdalla, W. Zhu, L. Xie, Y. Wang, and X. Guo, "Systematic comparison of two animal-to-human transmitted human coronaviruses: SARS-CoV-2 and SARS-CoV," Viruses 12, 244. 2020.

[18] G.M. Leung, L.M. Ho, T.H. Lam, and A.J. Hedley, "Epidemiology of SARS in the 2003 Hong Kong epidemic," Hong Kong Med. J. 15, 12–16. 2009.

[19] R. Viner, S. Russell, H. Croker, J. Packer, J. Ward, C. Stansfield, O. Mytton, and R. Booy, "School Closure and Management Practices during Coronavirus Outbreaks including COVID-19: A Rapid Narrative Systematic Review," Available at SSRN 3556648. 2020.

[20] M. Yu, Z. Li, Z. Yu, J. He, and J. Zhou, "Communication related health crisis on social media: a case of COVID-19 outbreak," Current Issues in Tourism, pp.1-7. 2020.