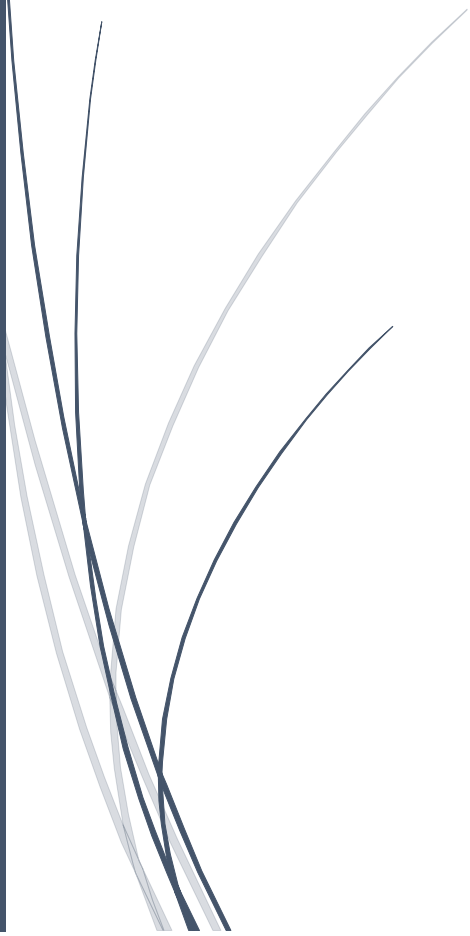
BIG DATA MANAGEMENT

10/20/2023

CSC 6002 FINAL PROJECT REPORT

Student Name: Chukwunonso Noella Akaeme



SEMESTER 2, 2023

UNIVERSITY OF SOUTHERN QUEENSLAND

Table of Contents

1 INTRODUCTION

Mobility Dataset Details: The dataset contains a global mobility report from over 100 countries, including Australia, during the COVID-19 pandemic. It includes over a million cases featuring geographical locations, metrics, and sub-regions. The data shows percentage changes in human movement compared to 2020 as a baseline period.

Covid Case Dataset Details: The dataset contains information about COVID-19 cases in 190 countries, including Australia. It includes over 50,000 cases featuring, dates, deaths, vaccine information, human development index and the likes. Overall, it provides a comprehensive view of the pandemic's impact on the population and the country's development within 2020 and 2021.

Where Mobility dataset was downloaded: I downloaded the mobility dataset from the *Assessment resources* under the *assessment* section of the CSC 6002 course homepage, the csv file was attached. Alternatively, there was also a download link from the Data Source: < https://www.google.com/covid19/mobility/>.

2 DATA EXPLORATION

Size of Mobility Dataset: The dataset is large containing a total of 135 countries with over 1 million cases which means that each record represents a specific mobility report for a particular location in a country and a specific date.

Size and Format of the Covid Case Dataset: the dataset is relatively large with a total of 62,407 cases from 190 countries including world and international records. There is a lot of information provided in the dataset about the COVID-19 situation across countries, which makes it valuable for in-depth analysis, modelling, and research focusing on the pandemic's impact on society and healthcare. The dataset is in CSV (Comma-Separated Values) format, organized into columns and rows containing text, numeric values and dates as per the data type.

Format of Mobility Dataset: The dataset is in CSV (Comma-Separated Values) format, organized into columns and rows containing text, numeric values and dates as per the data type. There also happens to be some missing values with the numeric data types such as the given percentage changes, codes and unique identifiers. Some countries also have missing sub-regions.

Features of the Mobility Dataset: The dataset contains 15 columns, each representing a specific feature of mobility reports. The first five columns provide location information, while the last six columns show percentage changes, retail, recreation, grocery, pharmacy, outdoor activities, public transportation usage, work-related mobility, and home stays compared to 2020.

Features of the Covid Case Dataset: The dataset contains 59 columns, each representing a specific feature of the Covid-19 cases in various countries. The first four columns provide basic information about location and dates, the next twelve columns provide information about the COVID-19 cases and death statistics, the next nine columns provide details about healthcare and the reproduction rate of the virus, the nine columns that followed provides information about COVID tests conducted and their results, the next nine columns featured vaccination information and the last 16 columns provide information about demographic, economic and healthcare factors, each of these elements could have played a role in the expansion and effects of the pandemic.

3 LITERATURE REVIEW

Research Works Featuring the Mobility Dataset:

- *Using A Partial Differential Equation with Google Mobility Data to Predict COVID-19 in Arizona* by Wang, H., & Yamamoto, N. (2020). Published by July 28, 2020.
- *Stay-at-Home Works to Fight Against COVID-19: International Evidence from Google Mobility Data* by Yilmazkuday, H. (2020). Published by November 1, 2020.

How The Aforementioned Works Answer Their Posed Research Question:

- The research paper "Stay-at-Home Works to Fight Against COVID-19: International Evidence from Google Mobility Data" by Hakan Yilmazkuday investigates the causal relationship between changes in mobility and COVID-19 cases/deaths in 130 countries using Google mobility data, citing other research works (Chan et al., 2020; Kraemer et al., 2020; Yilmazkuday, 2020) that demonstrates the efficiency of control measures and mobility restrictions in battling COVID-19 in particular nations but this paper stands out for achieving international comparisons using google mobility data. It was done by adjusting for variables like country-fixed effects, day-fixed effects, and country-specific timing of the 100th COVID-19 case, the research methodology employs a difference-in-difference design to regress weekly changes in COVID-19 cases and deaths on weekly changes in mobility measures. The empirical results suggest a clear correlation between mobility changes and COVID-19 outcomes. The study also finds that being at residential places has the most significant impact on reducing COVID-19 deaths compared to other mobility measures.
- In the research paper “Using A Partial Differential Equation with Google Mobility Data to Predict COVID-19 in Arizona” (Wang et al., 2020), the use of a spatio-temporal model based on partial differential equations (PDE) to predict the number of COVID-19 cases in the state of Arizona, USA was employed. The model makes use of the Google mobility data, reflecting human activities/interactions to understand the combined impact of the spread of the virus across regions and the behaviour of people regarding the transmission of COVID-19 (Wang et al.,2020). The study aims to understand the effectiveness of social distancing measures in minimizing the spread

of the virus by studying mobility in Arizona, the author explores how the removal of restrictions and the lack of the use of face masks by people led to the increase in the number of COVID-19 cases in the state. It also highlighted the importance of social distancing to prevent infections in the absence of a vaccine and the need to quantify the impact of such measures to convince the people and the government of their significance.

Research Works Featuring the Covid Cases Dataset:

- '*Clustering analysis of countries using the COVID-19 cases dataset*' by Zarikas, V, Poulopoulos, SG, Gareiou, Z & Zervas, E (2020). Published by 29 May 2020.
- '*COVID-19 Cases and Deaths in Southeast Asia Clustering using K-Means Algorithm'* by Hutagalung J, Ginantra NLWSR, Bhawika GW, Parwita WGS, Wanto A and Panjaitan PD (2021). Published by 2021.

How The Aforementioned Works Answer Their Posed Research Question:

- The research paper *"Clustering analysis of countries using the COVID-19 cases dataset"* (Zarikas et al.,2020), investigates how countries can be clustered based on the progression of COVID-19 cases, and the insights the clusters provide for understanding and managing of the pandemic. They did this by taking 30 nations with the oldest and highest COVID cases out of the Johns Hopkins University website and extracting the time series data of confirmed, recovered, and current cases then they developed a unique clustering technique that modifies the size, scope, and form of the time series and employs hierarchical clustering with Euclidean distance and single/complete linkage. The authors used this algorithm to cluster countries based on three variables: their active cases, active cases per population, and active cases per population and per area, resulting in different clusters reflecting the pandemic's impacts in different countries. Finally, they emphasized the importance of these clustering results for policymakers and researchers seeking to understand and manage the COVID-19 pandemic.
- In the research paper, "*COVID-19 Cases and Deaths in Southeast Asia Clustering using K-Means Algorithm"* (Hutagalung J et al., 2021), k-means algorithm was employed to classify COVID-19 cases and deaths in South-east Asia. The researchers utilised WHO data on the total number of confirmed cases and fatalities from COVID-19 in 11 South-east Asian nations from April 1 to April 30, 2020. They used K-means clustering approach with RapidMiner tool to split the data into three clusters: high (C1), medium (C2), and low(C3). To calculate the initial centroid values, they used the maximum, mean, and lowest values of the attribute values for each cluster and iterated until the centroid values were ideal then they displayed the clustering findings in tables, graphs, and maps. Additionally, the researchers, analysed the results and addressed the consequences for each country in terms of COVID-19 transmission and prevention.

4 RESEARCH QUESTION/SELECTION OF THE PROBLEM

4.1 Research question

"How did the rise in COVID-19 cases in 2020 change the way people moved around in Australia and the UK?"

This question will allow us to explore how the increase in total COVID-19 cases might have influenced people's behavior in terms of mobility in these two countries.

4.2 Justification of research question

- **Public Health Impact:** By examining changes in mobility patterns in conjunction with COVID-19 case counts, we may learn more about how people's movement changes in response to the pandemic. This can aid in our understanding of how well public health policies and recommendations work and the manner in which both nations successfully contained the outbreak.
- **Comparative Analysis:** By focusing on Australia and the UK, we can compare and contrast the situation in these two countries, which have different geographical locations, population densities, and healthcare systems. This could potentially reveal strategies or measures that have been particularly effective in one country but not in the other.
- **Policy Analysis:** The research question allows for an examination of the procedures and steps taken by Australia and the UK. It contributes to understanding whether more rigorous actions or softer tactics had a superior influence in curbing the COVID-19 spread, which could be beneficial for future policy formulation.
- **Future Pandemic Preparedness:** Identifying the relationship between mobility and covid cases will also help future pandemic preparedness, helping Australia and the UK respond more effectively to similar cases.

4.3 Importance of comparing COVID-19 situation between Australia and other countries

The research question is important for comparing the COVID-19 situation between Australia and other countries because:

- It can help to understand how people's behaviour and mobility patterns have changed in response to the pandemic and the public health measures implemented by different governments.
- It can help to evaluate the effectiveness and impact of different lockdown strategies and social distancing policies on reducing the spread of the virus and protecting the health of the population.

- It can help to identify the areas and sectors that have been most affected by the mobility restrictions and the potential economic and social consequences of the pandemic.
- It can help to inform future policy decisions and interventions to mitigate the risks and challenges posed by COVID-19 and other emerging infectious diseases.

5 METHODS

5.1 Definition of Spark

Spark is an open-source distributed computing framework that supports a multitude of programming languages, SQL, machine learning, graph analysis, and streaming, and can run on a wide range of platforms.

5.2 Reason for choosing Spark over Hadoop MapReduce

Spark was selected over Hadoop MapReduce due to its superior speed, versatility, and user-friendliness. Spark can perform in-memory processing, which minimises the disk I/O and network overhead associated with MapReduce. Furthermore, Spark can manage a variety of data sources and formats, including structured, semi-structured, and unstructured data. Spark's extensive collection of APIs and libraries simplifies the development and deployment of complex applications. Additionally, Spark has a user-friendly interactive shell that can facilitate data exploration and analysis.

6 CONNECTION BETWEEN THE DATASETS

Connection of both datasets to answer the research question:

To answer the research question, "How did the rise in COVID-19 cases in 2020 change the way people moved around in Australia and the UK?" using the two datasets, these steps need to be executed:

- Clean both datasets, removing duplicates and handling missing values.
- Integrate the datasets by finding common attributes like 'location' and 'date'.
- Subset the data for Australia and the UK, making the analysis more manageable.
- Calculate the correlation between total cases and mobility features.
- Visualize the results using graphs and charts.
- Draw conclusions based on the analysis.

How to select relevant subsets of the datasets:

To find useful subsets in the datasets, we employed the following steps:

- Install java
- Install spark
- Create spark session
- Import the datasets for COVID-19 cases and the global mobility report.
- Investigate the structure and features of both datasets.

- Find a shared identifier for merging, such as “date” and “iso_code”
- Combine the datasets using the shared identifier.
- Refine the combined dataset to only include data for Australia and the UK.
- Limit the analysis to the relevant years, which are 2020 and 2021.
- Select the necessary fields for the analysis, including date, mobility-related fields, and COVID-19-related fields.

7 DATA ANALYSIS

7.1 Filtering and Merging Datasets using Spark

We installed java and Spark using python programming language on Google Colaboratory with appropriate commands and code to filter, analyse and merge the datasets before converting to *pandas* dataframe for further analysis and visualisation. The Spark code was renamed *“spark_code”* and was submitted alongside this report as required.

7.2 Visualisations and Analysis

This subsection features the analysis and visualisations of the connected covid cases and mobility datasets to answer our research question:

Correlation between the Mobility trends and Covid cases in Australia:

Table 1: correlation for Australia

CORRELATION FOR AUSTRALIA	
Total Cases and Retail and Recreation Mobility	0.21
Total cases and Grocery and Pharmacy Mobility	0.015
Total Cases and Mobility to Parks	0.13
Total Cases and Transit Stations Mobility	-0.047
Total Cases and Workplace Mobility	-0.009
Total Cases and Residential Mobility	-0.012

Correlation heatmap for Australia:

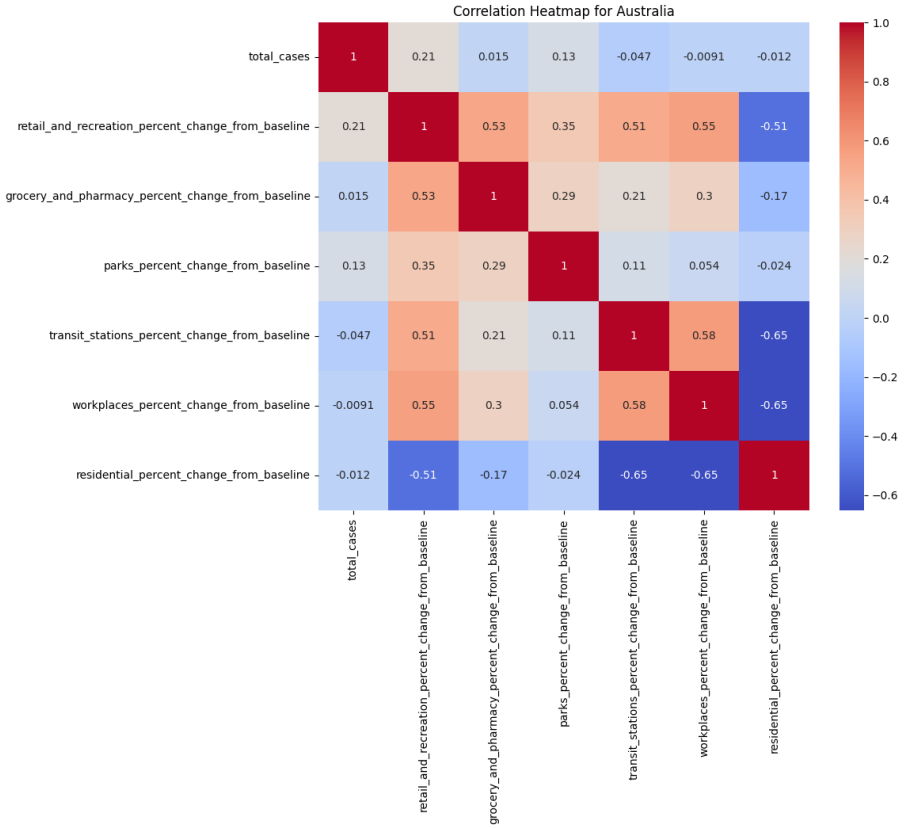


fig 1: correlation heatmap for Australia

Discussion of Correlation result for Australia: The correlation values indicate the relationship between total COVID-19 cases and various mobility trends in Australia.

- The correlation between total cases and retail and recreation is 0.207, indicating a weak positive relationship. This suggests that as COVID-19 cases increased, visits to retail and recreation places also slightly increased.
- The correlation with grocery and pharmacy is very weak (0.015), suggesting almost no relationship with total cases.
- The correlation with parks is weakly positive (0.128), indicating a slight increase in park visits as cases increased.
- The correlations with transit stations, workplaces, and residential areas are negative but very weak, suggesting almost no relationship with total cases.

Overall, the correlation results indicates a weak positive relationship between the total number of COVID-19 cases and the frequency of visits to retail and recreational facilities, grocery stores and pharmacies, and parks. This suggests that as COVID-19 cases rose, Australians visited these places more often. This could be due to Australia's relatively lower incidence of Covid cases and more lenient lockdown restrictions compared to other nations. On the other hand, there's a weak negative association between the total cases and the usage of transit stations and mobility to workplaces. This suggests that as the COVID-19 cases rose,

there was a decrease in public transport usage and workplace attendance. This could be because individuals opted for different commuting options such as their own private cars or chose to work remotely to lower their risk of infection. There was no significant association between the total cases and residential mobility trends, suggesting that the number of COVID-19 cases didn't have a substantial impact on the time people spent at home.

Correlation between the Mobility trends and Covid cases in the UK:

Table 2: correlation for the UK

CORRELATION FOR THE UK	
Total Cases and Retail and Recreation Mobility	0.0042
Total cases and Grocery and Pharmacy Mobility	0.13
Total Cases and Mobility to Parks	-0.08
Total Cases and Transit Stations Mobility	-0.18
Total Cases and Workplace Mobility	-0.006
Total Cases and Residential Mobility	0.11

Correlation heatmap for the UK:

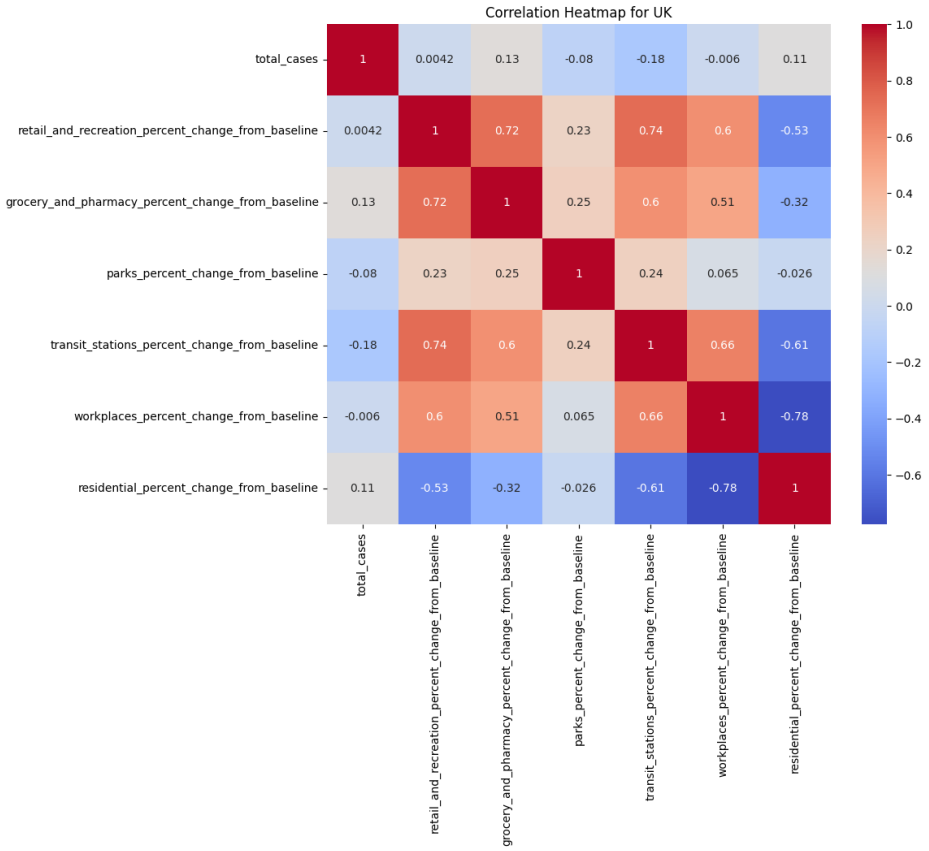


fig 2: correlation heatmap for UK

Discussion of Correlation result for the UK:

- The correlation between total cases and retail and recreation is very weak (0.004), suggesting almost no relationship.
- The correlation with grocery and pharmacy is weakly positive (0.127), indicating a slight increase in visits as cases increased.
- The correlation with parks is weakly negative (-0.080), suggesting a slight decrease in park visits as cases increased.
- The correlation with transit stations is weakly negative (-0.176), indicating a slight decrease in transit station visits as cases increased.
- The correlations with workplaces and residential areas are very weak, suggesting almost no relationship with total cases.

Overall, In the United Kingdom, the correlation results indicate a weak to moderate negative association between the total number of COVID-19 cases and all mobility trends, excluding residential mobility. This suggests that as the COVID-19 cases rose, people in the UK visited retail and recreational areas, grocery stores and pharmacies, parks, transit stations, and workplaces less frequently. This could be due to the UK's high covid case count and the implementation of strict lockdown measures to control the virus spread. There's a weak positive correlation between the total cases and residential mobility trend. This implies that as the COVID-19 cases escalated, people spent more time at home. This could be because individuals adhered to stay-at-home orders and refrained from going out unless absolutely necessary.

Timeseries of Retail and Recreation Mobility/Total Cases in Australia:

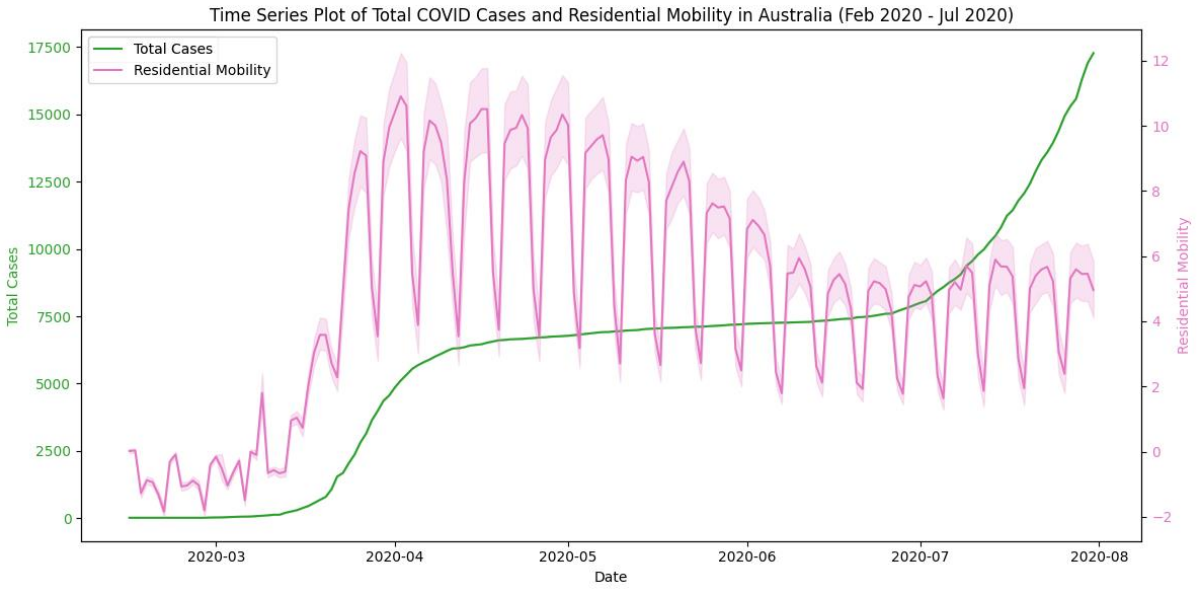


Fig 3: time series for residential mobility and total cases in Australia

Discussion of Timeseries plot of Retail and Recreation Mobility/Total Cases in Australia:

Based on the correlation analysis, we plotted time series for the mobility trends that have the strongest absolute correlation with total cases for each country. This is because these trends are most strongly associated with changes in COVID-19 cases and visualising them would provide valuable insights into how people's movements changed in response to the increase in covid cases in 2020.

For Australia, the mobility trend with the strongest absolute correlation is *`retail_and_recreation_percent_change_from_baseline`*. We thought that since there is a weak positive correlation (0.207), it might be interesting to see how these two trends move together over time.

From the timeseries, the green line represents the total cases of COVID-19 in Australia from February 2020 to July 2020. It shows a steep increase from February to April, indicating a rapid rise in COVID-19 cases during this period. From May to July, the line gradually increases, suggesting that the number of cases continued to rise, but at a slower rate but skyrocketed at the end of July to 17,500 cases.

The pink line represents residential mobility. It shows a sharp decrease from February to March, which could be due to lockdown measures and restrictions put in place to control the spread of the virus. From April to July, the line gradually increases, indicating that residential mobility started to improve as restrictions may have been eased.

Timeseries of Grocery and Pharmacy Mobility/Total Cases in the UK:

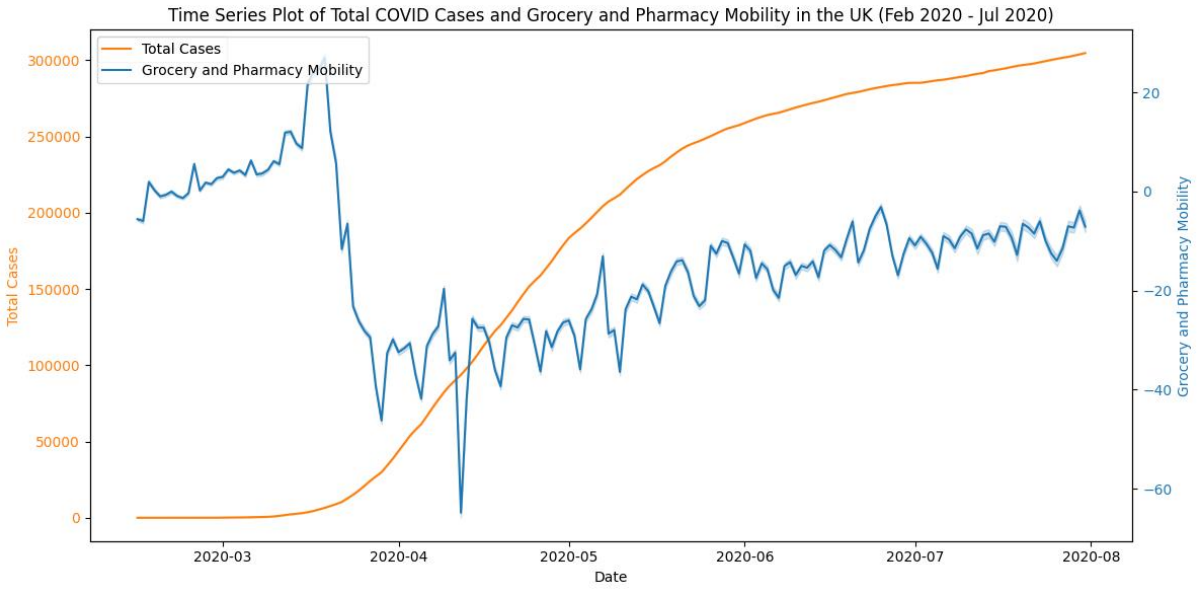


Fig 4: time series for grocery/pharmacy mobility and total cases in the UK

Discussion of Timeseries plot of Grocery/Pharmacy Mobility and the Total Covid Cases in the UK:

For the UK, we plotted '*Total Cases and Grocery & Pharmacy'*. since there is a weak positive correlation (0.127), and *'Total Cases and Transit Stations'*: since there is a weak negative correlation (-0.176), indicating that as cases increased, visits to grocery stores and pharmacies also slightly increased while visits to transit stations slightly decreased as cases increased.

The time series plot shows the total COVID-19 cases and grocery and pharmacy mobility in the UK from February 2020 to July 2020. The blue line represents grocery and pharmacy mobility. It shows a decrease in mobility from February to April with April coming down to a -60% change in mobility, which could be due to lockdown measures and restrictions put in place to control the spread of the virus. From April to July, the line gradually increases, indicating that grocery and pharmacy mobility started to improve as restrictions may have been eased. The gradual increase in grocery and pharmacy mobility could also indicate that people were adapting to new ways of shopping, such as online ordering during this period.

The orange line in the graph represents the total number of COVID-19 cases in the UK from February 2020 to July 2020. It shows a sharp increase in cases from February to April, indicating a rapid rise in COVID-19 cases during this period. From April to July, the line gradually increases, suggesting that the number of cases started to increase. This could be due to a variety of factors, such as the non-adherence to public health measures, or changes in individual behavior in response to the pandemic.

Timeseries of Transit stations Mobility/Total Cases in the UK:

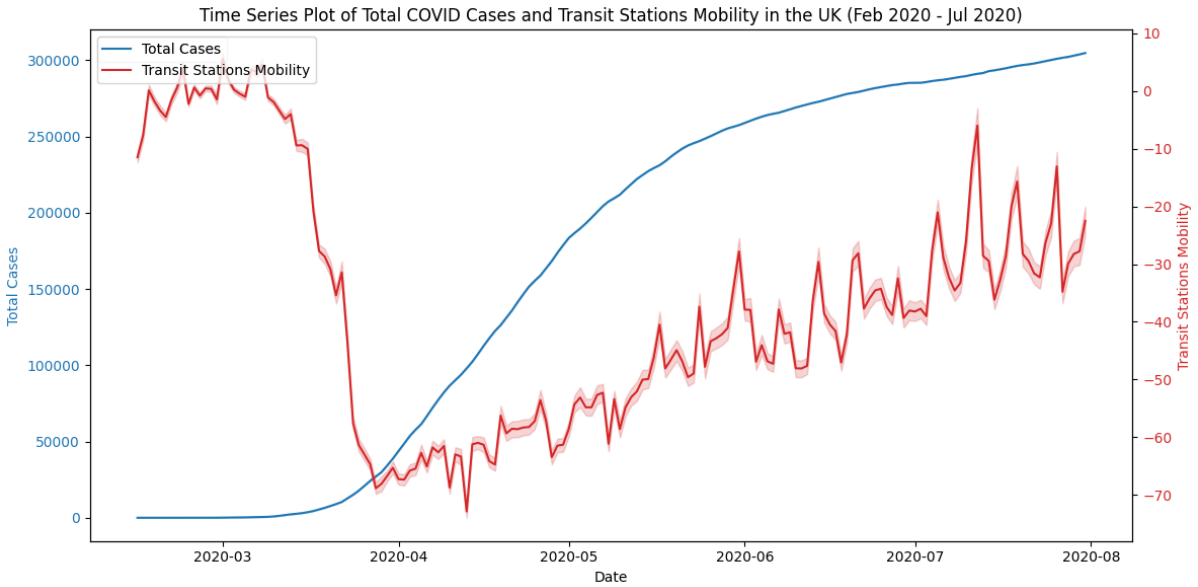


Fig 5: time series for transit stations mobility and total cases in the UK

Discussion of Timeseries plot of Transit stations Mobility and the Total Covid Cases in the UK:

The figure above is a time series plot of total COVID-19 cases and transit station mobility in the UK from February 2020 to July 2020.

The blue line represents the total number of COVID-19 cases. It shows a steady increase in cases from February to July, indicating that the pandemic was continuously spreading during this period.

The red line represents transit station mobility. It shows a decrease in mobility from February to April, which could be due to lockdown measures and restrictions put in place to control the spread of the virus. From April to July, the line shows a slight increase, indicating that transit station mobility started to improve as restrictions may have been eased allowing people in the UK to commute via public transits.

8 FINDINGS

Our research question is: ***How did the rise in COVID-19 cases in 2020 change the way people moved around in Australia and the UK?***

Discussion to answer research question based on the findings from the analysis:

- The rise in COVID-19 cases in 2020 had a significant impact on the way people moved around in Australia and the UK. Both countries experienced a decrease in mobility across different sectors, such as retail, recreation, transit stations, and workplaces, as lockdown measures and restrictions were implemented to control the spread of the virus.
- However, there were also some differences between the two countries. Australia had a lower number of cases and a higher residential mobility than the UK, indicating that people stayed more at home and followed public health guidelines. The UK had a higher number of cases and a higher grocery and pharmacy mobility than Australia, suggesting that people may have increased their visits to essential shops or adapted to online shopping during the pandemic.
- The correlation analysis and the time series plots showed that some mobility trends were more strongly associated with changes in COVID-19 cases than others. For Australia, retail and recreation mobility had a weak positive correlation with total cases, while for the UK, grocery and pharmacy mobility and transit station mobility had weak positive and negative correlations respectively with total cases. These trends reflected how people’s movements changed in response to the increase in COVID-19 cases in 2020.

It is important to note that our conclusions are based on correlation coefficients. Correlation does not equate to causation, and there may be other influencing factors not accounted for in this data. While there is a correlation between these mobility trends and the covid case numbers, it doesn't necessarily mean that changes in case numbers directly caused these

mobility changes. Also, people's reactions to the pandemic are multifaceted and can be shaped by numerous factors beyond just the number of cases, such as government policies, public health communications, and individual beliefs and attitudes.

9 ETHICS AND PRIVACY

Australian law on collecting public data/Validity of Mobility dataset:

The Privacy Act of 1988 governs the gathering of personal information in Australia. This legislation states that an organisation may only gather personal data that is legitimately required for them to do their jobs (OAIC, n.d.). Under the Australian law, individuals generally have the right to provide or withhold consent for the collection of their private information and the organizations are encouraged to anonymize data whenever possible.

According to the Australian law, this community mobility dataset is valid as it adheres to the guidelines outlined in the Privacy Act of 1988. The data set have anonymized the identity of the individuals to reduce privacy risks as there are no names, address or personal information pertaining to the individuals, all information was generalized and public based. It also was solely and legitimately for research purposes pertaining to the COVID-19 spread and since the consumers gave approval, they have complete control over what information is shared and how it may be utilised given that the Consumer Data Right (CDR) system will only allow for a consumer's data to be shared with providers they have chosen.

Australian law on collecting public data/Validity of Covid case dataset:

The Privacy Act of 1988 governs how personal information is collected and used in Australia, by the use of the surface principles known as “APPs” which stands for the Australian Privacy Principles (OAIC, n.d.; GoCardless, n.d.; Ground Labs, 2021). Under the Australian law, individuals generally have the right to provide or withhold consent for the collection of their private information and the organizations are encouraged to anonymize data whenever possible. The information should also be obtained directly from the individual rather than through a third party and must be gathered lawfully and fairly (OAIC, n.d.).

The COVID-19 case dataset is valid according to the Australian legislation. This is due to the lack of any apparent personal information about specific persons, instead, it presents compiled information on COVID-19 case counts throughout a number of countries, including Australia. The typical reasons for collecting and using this kind data are research and public health monitoring.

10 HOSTING ON A SERVER

I have attached below, the video link showing the running of the spark code in the Jupyter notebook server on the AZURE virtual machine.

https://drive.google.com/file/d/1azRLis84O4yx3z3ikt9vY74Z4r9nkILL/view?usp=sharing

11 REFERENCES

1. Mathieu, E., Ritchie, H., Rodés-Guirao, L., Appel, C., Giattino, C., Hasell, J., Macdonald, B., Dattani, S., Beltekian, D., Ortiz-Ospina, E., and Roser, M., 2020. Coronavirus pandemic (COVID-19). [online] Our World in Data. Available at:<https://ourworldindata.org/coronavirus > [Accessed 13 Oct. 2023].
2. Wang, H., & Yamamoto, N. 2020. *Using A Partial Differential Equation with Google Mobility Data to Predict COVID-19 in Arizona*.
3. Yilmazkuday, H. 2020. *Stay-at-Home Works to Fight Against COVID-19: International Evidence from Google Mobility Data*.
4. Murnane, K., 2019. *Three steps google is taking to enhance user's data privacy*. [online] Forbes. Available at: <https://www.forbes.com/sites/kevinmurnane/2019/05/11/three-steps-google-is-taking-to-enhance-users-data-privacy/?sh=71fb847a163e.> [Accessed 13 Oct. 2023].
5. Zarikas, V, Poulopoulos, SG, Gareiou, Z & Zervas, E 2020, '*Clustering analysis of countries using the COVID-19 cases dataset*', Data in Brief, vol. 31, p. 105787, viewed 13 October 2023, doi:10.1016/j.dib.2020.105787.
6. Hutagalung J, Ginantra NLWSR, Bhawika GW, Parwita WGS, Wanto A and Panjaitan PD 2021, *COVID-19 Cases and Deaths in Southeast Asia Clustering using K-Means Algorithm*, Journal of Physics: Conference Series, 1783, 012027, viewed 13 October 2023, doi:10.1088/1742-6596/1783/1/012027.
7. Oaic, 2023. *Guide to data analytics and the Australian privacy principles*. [online] OAIC. Available at: <https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/more-guidance/guide-to-data-analytics-and-the-australian-privacy-principles> [Accessed 14 Oct. 2023].
8. Oaic, 2023a. *Collection of personal information*. [online] OAIC. Available at: <https://www.oaic.gov.au/privacy/your-privacy-rights/your-personal-information/collection-of-personal-information> [Accessed 14 Oct. 2023].