Justin Ross

0050040865

CSC8002 – Big Data Management

Assignment 1

Contents

Introduction

Discussion

The dataset being used is Google's COVID-19 Community Mobility Reports, which is a record of changes in people's movement, grouped by usage of specific area types such as residential, park, etc. Google has created this dataset through two processes. First the data is obtained from users who have their location history turned on in their Google account and who are using mobile devices. In other words the location history tracks the user's location via their smart phone. Secondly, the report uses a baseline day as a benchmark to compare area usage and then records the change relative to the benchmark. The baseline day is a median value from like days comprising the 5 week period between 3/01/2020 and 06/02/2020[1]. That is to say Fridays are compared to the median value from the 5 Fridays between the 03/01/2020 and the 06/02/2020, which is the baseline.

Source

The dataset was downloaded from Google: https://www.google.com/covid19/mobility/, though access has been granted only temporarily by Google to help with the coronavirus epidemic.

Usefulness of the Dataset

The dataset can help Australia to tackle COVID 19 through a number of mechanisms. Firstly, it can identify areas where there are a lot of people moving about and that are therefore risk areas or areas where people are not adhering to distancing rules. Secondly, it can identify the association between mobility in an area and the spread of COVID 19 should an outbreak occur, thus identifying how effective management techniques proved to be.

[1] (Aktay et al. 2020, p. 4-5)

Data Exploration

Size

The dataset contains 3342175 entries or rows, with 14 columns and has a file size of 223 MB. It is worth noting that this may be subject to change as the data from Google is dynamic and updated over time.

Format

The format of the dataset is comma separated variables (CSV), which means that it is basically a text file with entries in each row being separated by commas per column value. Overall the columns fall roughly into three categories: geographic data to give the location, the change in use of specific area types relative to the baseline and the date.

Features

It has 14 columns with varying content as detailed in the table below:

Column Name	Content
country_region_code	Country Code, IE: AU
country_region	Country Name, IE.: Australia
sub_region_1	State
sub_region_2	City
metro_area	Non-Australia Suburb Designation
iso_3166_2_code	State in code form, IE: AU-ACT
census_fips_code	US specific geographic coding
Date	The date the stats were recorded
retail_and_recreation_percent_change_from_baseline	Optional shopping and recreation
grocery_and_pharmacy_percent_change_from_baseline	Essential shopping
parks_percent_change_from_baseline	Parks, national parks, castles, etc.
transit_stations_percent_change_from_baseline	Train stations, bus stations, etc
workplaces_percent_change_from_baseline	Work sites, corporate areas.
residential_percent_change_from_baseline	Homes, apartments, etc.

Table 1: A table of the features/columns and the content they contain.

All of the features that include the word ‘baseline’ basically function to measure the change in use of an area compared to the median of the baseline days before the period being measured. For instance, retail_and_recreation_percent_change_from_baseline will have negative or positive values indicating the percentage change from the median of the baseline days.

Literature Review

Research Example 1 - COVID-19 outbreak response, a dataset to assess mobility changes in Italy following national lockdown

The first piece of research I have chosen is ‘COVID-19 outbreak response, a dataset to assess mobility changes in Italy following national lockdown’ by Emanuele Pepe and team. The article documents the teams work focusing on establishing just how effective non-pharmaceutical interventions (NPIs) have been at controlling the spread of COVID-19 in Italy[2]. It uses the mobility data from Google to examine whether measures put in place by governmental authorities have proven successful in changing the movement patterns of the population.

This paper takes quite a sophisticated approach to using the data, feeding it into Gyration radii to examine how much area a person inhabited and comparing it to others. The idea appears to be to model the movement of people and assess risk based on number of ‘collisions’ between different people and therefore the risk of virus spread[3].

Research Example 2 - Interpreting the effect of social restrictions on cases of COVID-19 using mobility data

The second piece of research I have chosen is ‘Interpreting the effect of social restrictions on cases of COVID-19 using mobility data’, which is a pre-print from the Medical Journal of Australia. This piece of research is particularly useful as it gives a very clear example of how the data set could be used to answer questions in the Australian context.

The researchers coupled data from Australian mobility trends with confirmed COVID-19 case data from John Hopkins University[4]. This involved adjusting for the lag period of 11 days between exposure and confirmation of a COVID-19 case. After which mobility trends were examined to check how effective social distancing techniques were in decreasing the ‘doubling time’, a measure of how quickly virus cases grow by measuring the time it takes for number of infected people to double[5].

[2] (Pepe et al. 2020, p.1)
[3] (Pepe et al. 2020, pp. 3-4)
[4] (Tran et al. 2020, p. 2)
[5] (Tran et al. 2020, pp. 3-4)

Research Question/Selection of the Problem

Selected Research Question

The selected research question is: ‘How effective were government social distancing measures in decreasing people using public areas?’. To approach this question change in use of non-public areas (residential and perhaps grocery) will be compared to all of the other areas.

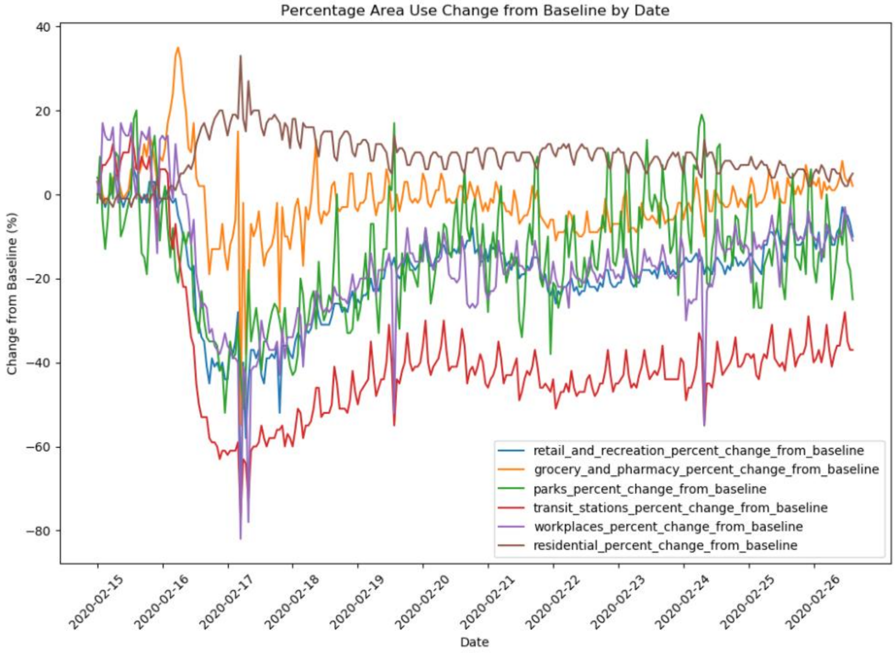

Figure 1: Change in area use over time compared to baseline.

Justification

Although the government put into place social distancing measures, the degree to which they were effective is not clear. Anecdotal measures may be used to assess effectiveness but these measures may not be rigorous. The Google Covid-19 mobility data allows the government and health community the ability to see how effective the measures were with solid, quantitative data. This means a balance can be struck between lockdown policies and public freedom as measures that prove ineffective at slowing the spread of the virus may be exempted[6].

The specific question chosen is valuable because it correlates the lockdowns and social distancing measures with data of people’s movements. It also does not need much in the way of further data save for the implementation dates of policies after which the change in behavior of people in a region such as Brisbane can be observed.

[6] (Tran et al. 2020, p. 3)

Privacy

Google employed a number of techniques to protect the privacy of its users. The first step was removing all information from the dataset that could potentially identify an individual such as name, address, etc. However, although this measure is a strong privacy measure and may be sufficient in many cases Google took further steps to protect user privacy.

The second step was to limit the publication of areas that were too small, with the minimum size for a geographic area being 3km^2 [7]. The idea with this privacy measure is that a small geographical area being discussed risks identifying individuals by their behavior, but if the area is large enough identifying individuals becomes impractical.

The third step was to remove observations in which the number of users were too small, namely when the number of users was less than 100[8]. The idea is that if the number of users is particularly small there is a risk that people will know which individual or small group was in a given area at a given time and will therefore be able to identify them. As such this data was excluded.

The fourth step was the employment of a more advanced mathematical technique, called 'differential privacy', which adds noise to the data without significantly affecting its accuracy[9]. The math here is somewhat sophisticated and involves running a query through a filter that applied 'Laplace noise' so that even the person extracting the data cannot know the exact details of it.

[7] (Aktay et al. 2020, p. 2)
[8] (Aktay et al. 2020, p. 3)
[9] (Aktay et al. 2020, p. 3)

Reference

Aktay, A., Bavadekar, S., Cossoul, G., Davis, J., Desfontaines, D., Fabrikant, A., Gabrilovich, E., Gadepalli, K., Gipson, B., Guevara, M., Kamath, C., Kansal, M., Lange, A., Mandayam, C., Oplinge, A., Pluntke, C., Roessler, T., Schlosberg, A., Shekel, T., Vispute, S., Vu, M., Wellenius, G., Williams, B. and Wilson, R., 2020. Google COVID-19 Community Mobility Reports: Anonymization Process Description (version 1.1). arXiv Pre-print,.

Pepe, E., Bajardi, P., Gauvin, L., Privitera, F., Lake, B., Cattuto, C. and Tizzoni, M., 2020. COVID-19 outbreak response, a dataset to assess mobility changes in Italy following national lockdown. Scientific Data, 7(1).

Tran, T., Sasikumar, S., Hennessy, A., O’Loughlin, A. and Morgan, L., 2020. Interpreting the effect of social restrictions on cases of COVID-19 using mobility data. The Medical Journal of Australia,.