Capstone Project - Final Assignment

COVID-19 Financial Impact & Venues Data Analysis (Part 1 and 2) by Alexander Nguyen

Applied Data Science Capstone by IBM/Coursera

June 2020

Table of contents

1. Introduction: Business Problem
2. Data
3. Methodology
4. Analysis
5. Results and Discussion
6. Conclusion

1. Introduction: Business Problem

The COVID-19 pandemic has had a devastating impact on a global scale, not only from a public health perspective but also with profound economic and financial consequences. In this project, I will attempt to perform an analysis of such financial impacts to selected suburbs in Melbourne, Australia where I currently live, based on their shared feature of most common venues in proximity.

This involves:

- Leveraging **Foursquare location data** and using the **k-means clustering algorithm** to segregate these suburbs into distinct groups;
- Adding **COVID-19 financial impact index** (publicly available data - more in ***Section 2. Data***) as an overlay; and
- Using **data visualisation** to draw insights and correlations to identify **which suburbs are most heavily impacted based on the predominant types of local businesses**.
While some inherent limitations of this analysis will also be discussed in ***Section 5. Results and Discussion***, it aims to showcase the application of data science in drawing useful insights to drive the decision making process for relevant stakeholders. For example, by showing that suburbs with a higher concentration of hospitality businesses (restaurants, coffee shops etc.) are more severely impacted financially than others, local and federal authorities may choose to direct more public funding to provide targeted support for these suburbs.

2. Data

In performing this analysis, I will use the following publicly available data sets. Sources have been quoted and all credits go to the creators. Results of this analysis are intended for training and demonstration purposes only.

a) Geospatial data (latitude and longitude) of inner suburbs in Melbourne, Australia
I've encountered issues with using the Google Maps Geocoding API. As such, I have used the publicly available **Free Database of Australia Postcodes** published at https://www.matthewproctor.com/full_australian_postcodes_vic (All credits go to Matthew Proctor).
b) Data of venues in proximity of the selected suburbs
Using **Foursquare API**, I obtained data of venues surrounding each suburb and their categories in order to segregate these suburbs into clusters based on their most common venues data.

c) COVID-19 Financial Impact Index
I used the **Taylor Fry COVID-19 Financial Impact Index** publicly available at https://taylorfry.com.au/articles/covid-19-financial-impact-index/ (All credits go to Taylor Fry).
This index provides "an estimate of financial impact – a substantial reduction in income relative to an individual's baseline expenses - by considering:

- Income – Whether an individual has lost significant income as a result of unemployment or underemployment, and to what extent government programs such as JobKeeper and JobSeeker cover pre-pandemic income.
- Expenditure – Whether current income is likely to cover non-discretionary expenses based on pre-pandemic income and life stage."

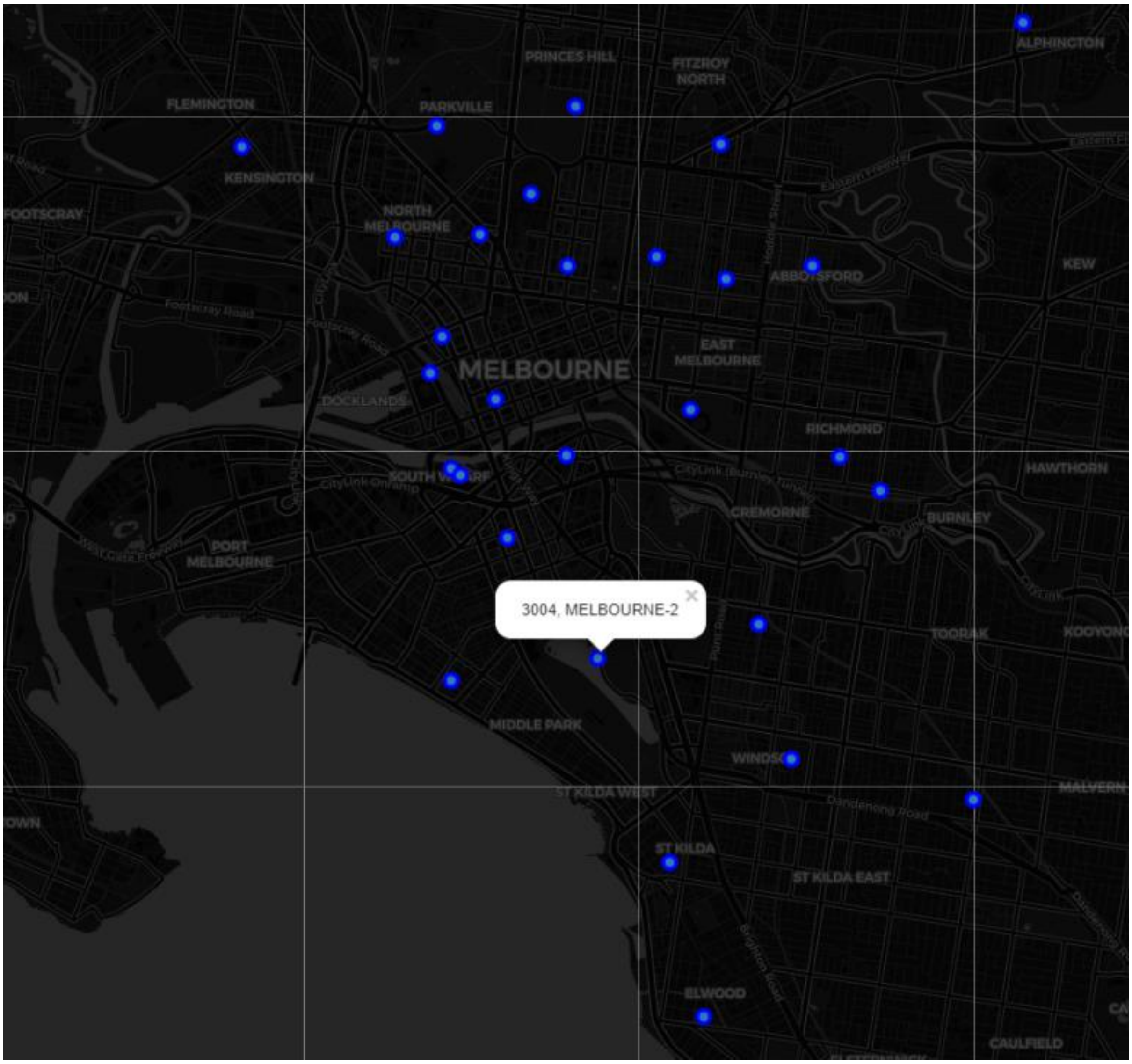
This data is presented by postcode in an interactive map format. For simplicity, once the relevant suburbs have been selected, I manually summarised the index data and merged it with the data set.

d) Median age by postcode 2016 Census Data
Demographic data added for additional analysis. Latest **2016 Census data** is publicly available from the Australian Bureau of Statistics (ABS). The data is presented by postcode in an interactive map format at https://quickstats.censusdata.abs.gov.au/census_services/getproduct/census/2016/quickstat/POA3000?opendocument. For simplicity, once the relevant suburbs have been selected, I manually summarised the Median Age data and merged it with the dataset.

Geospatial data extract

- For the purpose of this project, I will limit the scope of analysis to "Melbourne - Inner" suburbs only. The true postcodes should be limited to "Delivery Area" (Column "Category") and postcodes should be a 3000-series (Column "Postcode"), hence excluding for example those at the bottom of the data set.
- I will also drop some columns not required for this analysis (e.g. "ID", "Type").
- In this data set, multiple postcodes may have the same nominal longitude/latitude due to their proximity. For simplicity, I will keep only unique coordinates so they can be displayed on a map separately without being overlapped.

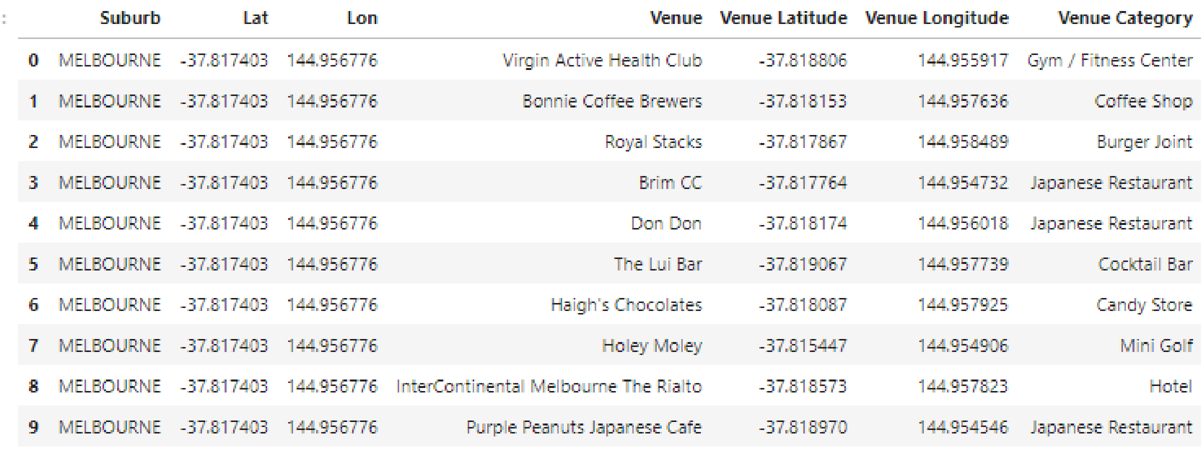
Once cleaned, we have **36 postcodes** in our data set. Let's now create a map to visualise this data set.



Foursquare

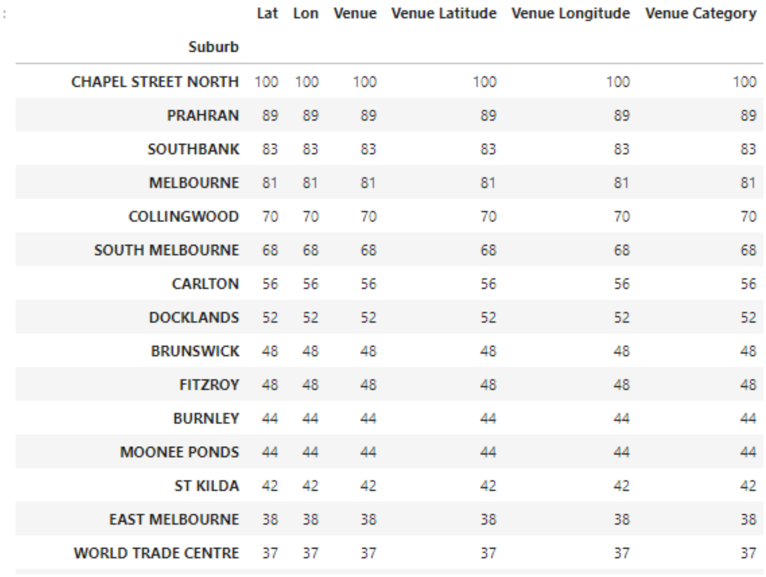
Having obtained a data set with central coordinates of the selected Melbourne inner suburbs, I will now use the Foursquare API to pull data of venues in proximity of each suburb.
I will limit the search to 100 venues within a radius of 500m from the defined coordinates.

As can be seen, the Foursquare search has returned **1223 venues** in **209 unique categories**.



	Suburb	Lat	Lon	Venue	Venue Latitude	Venue Longitude	Venue Category
0	MELBOURNE	-37.817403	144.956776	Virgin Active Health Club	-37.818806	144.955917	Gym / Fitness Center
1	MELBOURNE	-37.817403	144.956776	Bonnie Coffee Brewers	-37.818153	144.957636	Coffee Shop
2	MELBOURNE	-37.817403	144.956776	Royal Stacks	-37.817867	144.958489	Burger Joint
3	MELBOURNE	-37.817403	144.956776	Brim CC	-37.817764	144.954732	Japanese Restaurant
4	MELBOURNE	-37.817403	144.956776	Don Don	-37.818174	144.956018	Japanese Restaurant
5	MELBOURNE	-37.817403	144.956776	The Lui Bar	-37.819067	144.957739	Cocktail Bar
6	MELBOURNE	-37.817403	144.956776	Haigh's Chocolates	-37.818087	144.957925	Candy Store
7	MELBOURNE	-37.817403	144.956776	Holey Moley	-37.815447	144.954906	Mini Golf
8	MELBOURNE	-37.817403	144.956776	InterContinental Melbourne The Rialto	-37.818573	144.957823	Hotel
9	MELBOURNE	-37.817403	144.956776	Purple Peanuts Japanese Cafe	-37.818970	144.954546	Japanese Restaurant

Let's also check how many venues were returned for each suburb:



	Lat	Lon	Venue	Venue Latitude	Venue Longitude	Venue Category
Suburb						
CHAPEL STREET NORTH	100	100	100	100	100	100
PRAHRAN	89	89	89	89	89	89
SOUTHBANK	83	83	83	83	83	83
MELBOURNE	81	81	81	81	81	81
COLLINGWOOD	70	70	70	70	70	70
SOUTH MELBOURNE	68	68	68	68	68	68
CARLTON	56	56	56	56	56	56
DOCKLANDS	52	52	52	52	52	52
BRUNSWICK	48	48	48	48	48	48
FITZROY	48	48	48	48	48	48
BURNLEY	44	44	44	44	44	44
MOONEE PONDS	44	44	44	44	44	44
ST KILDA	42	42	42	42	42	42
EAST MELBOURNE	38	38	38	38	38	38
WORLD TRADE CENTRE	37	37	37	37	37	37

As can be seen, only "Chapel Street North" reached the 100 venues limit. The majority of suburbs returned fewer than 50 venues. This doesn't necessarily mean that this is an exhaustive list of results. In fact, it is affected by the use of only a single pair of coordinates as nominal location data for each suburb, as well as the radius used in the search query. Further implications to the analysis will be discussed in the next sections.

COVID-19 Financial Impact Index & Median Age data by Postcode

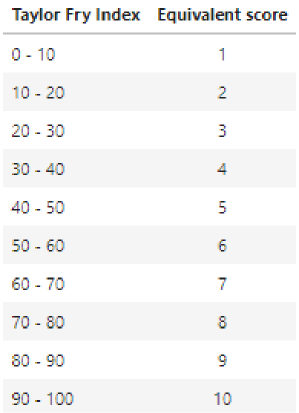
- As noted above, I will be using:

i) The Taylor Fry COVID-19 Financial Impact Index publicly available at https://taylorfry.com.au/articles/covid-19-financial-impact-index/; and
ii) 2016 Census data publicly available from the Australian Bureau of Statistics (ABS) at https://quickstats.censusdata.abs.gov.au/census_services/getproduct/census/2016/quickstat/POA3000?opendocument.
These data sets are presented in an interactive map format. For simplicity, I have manually summarised the data for the relevant suburbs into a data frame as below.

Note: The original Taylor Fry index is presented in a scale format such that:

- The 0 – 10 decile is the 10% of Australian postcodes least financially affected by COVID-19

- The 90 – 100 decile is the 10% of Australian postcodes most financially affected by COVID 19
  For simplicity, I have converted the index into equivalent scores as follows. Consistent with the index, the higher the score, the bigger the negative financial impact compared to pre-pandemic levels.

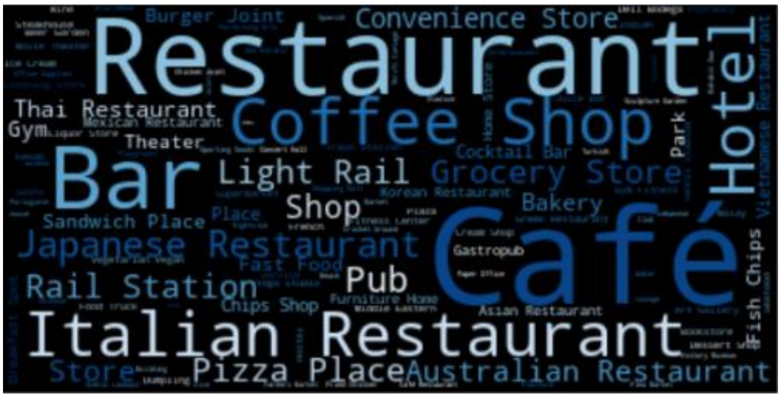


Taylor Fry Index	Equivalent score
0 - 10	1
10 - 20	2
20 - 30	3
30 - 40	4
40 - 50	5
50 - 60	6
60 - 70	7
70 - 80	8
80 - 90	9
90 - 100	10

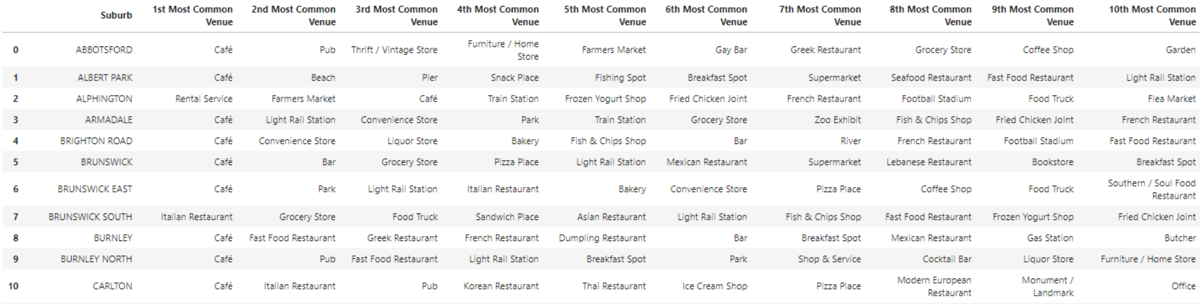
3. Methodology

- As noted above in ***Section 2. Data - Foursquare***, the Foursquare search query returned **1223 venues** in proximity of my selected Melbourne inner suburbs. It should be noted that this result is not exhaustive of the Foursquare venues data, as it depends on:
  i) The fact that I have used a pair of nominal central coordinates for each suburb. The result will vary with the accuracy of this input data. Suburb border geospatial data could also be used instead of just central coordinates, which may also give different results; and
  ii) Search limit of 100 venues within a radius of 500m. Increasing the limit and radius may give different results.

- What are the main venue categories? Let's try to visualise this:



- As I expected, Melbourne is called 'The coffee capital of the world' for a reason! Melbournians love their coffee, brunches and all sorts of cuisine. As such, we see the main types of venues such as 'Cafe', 'Restaurant', 'Coffee shop' and so on.
- I also expect a fair degree of homogeneity, i.e. many of Melbourne inner suburbs will have a high concentration of coffee shops and restaurants. As such, there may be similar financial impacts to these suburbs due to having similar types of business in proximity.
- Let's now drill down a bit further into the data set to see if my expectation is correct. I will show the most common types of venues by each suburb.



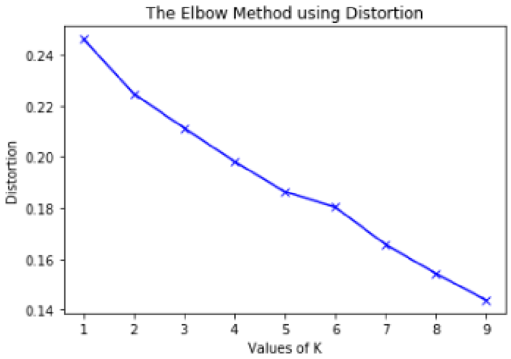
	Suburb	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	ABBOTSFORD	Café	Pub	Thrift / Vintage Store	Furniture / Home Store	Farmers Market	Gay Bar	Greek Restaurant	Grocery Store	Coffee Shop	Garden
1	ALBERT PARK	Café	Beach	Pier	Snack Place	Fishing Spot	Breakfast Spot	Supermarket	Seafood Restaurant	Fast Food Restaurant	Light Rail Station
2	ALPHINGTON	Rental Service	Farmers Market	Café	Train Station	Frozen Yogurt Shop	Fried Chicken Joint	French Restaurant	Football Stadium	Food Truck	Flea Market
3	ARMADALE	Café	Light Rail Station	Convenience Store	Park	Train Station	Grocery Store	Zoo Exhibit	Fish & Chips Shop	Fried Chicken Joint	French Restaurant
4	BRIGHTON ROAD	Café	Convenience Store	Liquor Store	Bakery	Fish & Chips Shop	Bar	River	French Restaurant	Football Stadium	Fast Food Restaurant
5	BRUNSWICK	Café	Bar	Grocery Store	Pizza Place	Light Rail Station	Mexican Restaurant	Supermarket	Lebanese Restaurant	Bookstore	Breakfast Spot
6	BRUNSWICK EAST	Café	Park	Light Rail Station	Italian Restaurant	Bakery	Convenience Store	Pizza Place	Coffee Shop	Food Truck	Southern / Soul Food Restaurant
7	BRUNSWICK SOUTH	Italian Restaurant	Grocery Store	Food Truck	Sandwich Place	Asian Restaurant	Light Rail Station	Fish & Chips Shop	Fast Food Restaurant	Frozen Yogurt Shop	Fried Chicken Joint
8	BURNLEY	Café	Fast Food Restaurant	Greek Restaurant	French Restaurant	Dumpling Restaurant	Bar	Breakfast Spot	Mexican Restaurant	Gas Station	Butcher
9	BURNLEY NORTH	Café	Pub	Fast Food Restaurant	Light Rail Station	Breakfast Spot	Park	Shop & Service	Cocktail Bar	Liquor Store	Furniture / Home Store
10	CARLTON	Café	Italian Restaurant	Pub	Korean Restaurant	Thai Restaurant	Ice Cream Shop	Pizza Place	Modern European Restaurant	Monument / Landmark	Office

As can be seen, the majority of suburbs have 'Cafe' as the 1st most common venue type! 'Bar', 'Pub' and other 'Restaurant' (Greek, Italian, Indian etc.) come very close in popularity as well.

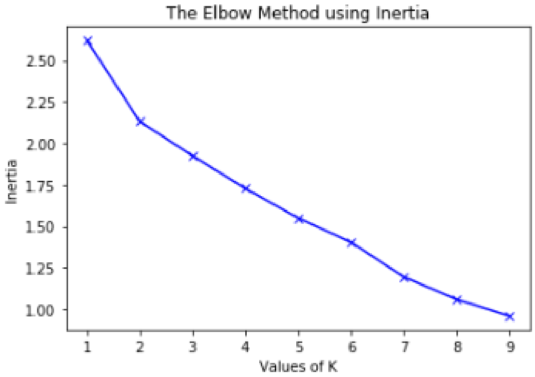
k-means clustering

- Next, I will attempt to use the k-means algorithm to group these suburbs into clusters based on these most common venues characteristics. K-means is considered one of the simplest and most commonly used unsupervised clustering methods. As such I will use k-means for simplicity and efficiency.
- As the data set is categorical in nature, to be able to use k-means I will use one-hot encoding to convert the attributes into dummy variables as follows.

- One of the key parameters used in k-means clustering is the targeted number of clusters. To determine the optimal number of clusters, we select the value of k at the "elbow" i.e. the point after which the distortion/inertia start decreasing in a linear fashion.
  i) Distortion is calculated as the average of the squared distances from the cluster centres of the respective clusters. Typically, the Euclidean distance metric is used.
  ii) Inertia is the sum of squared distances of samples to their closest cluster centre.

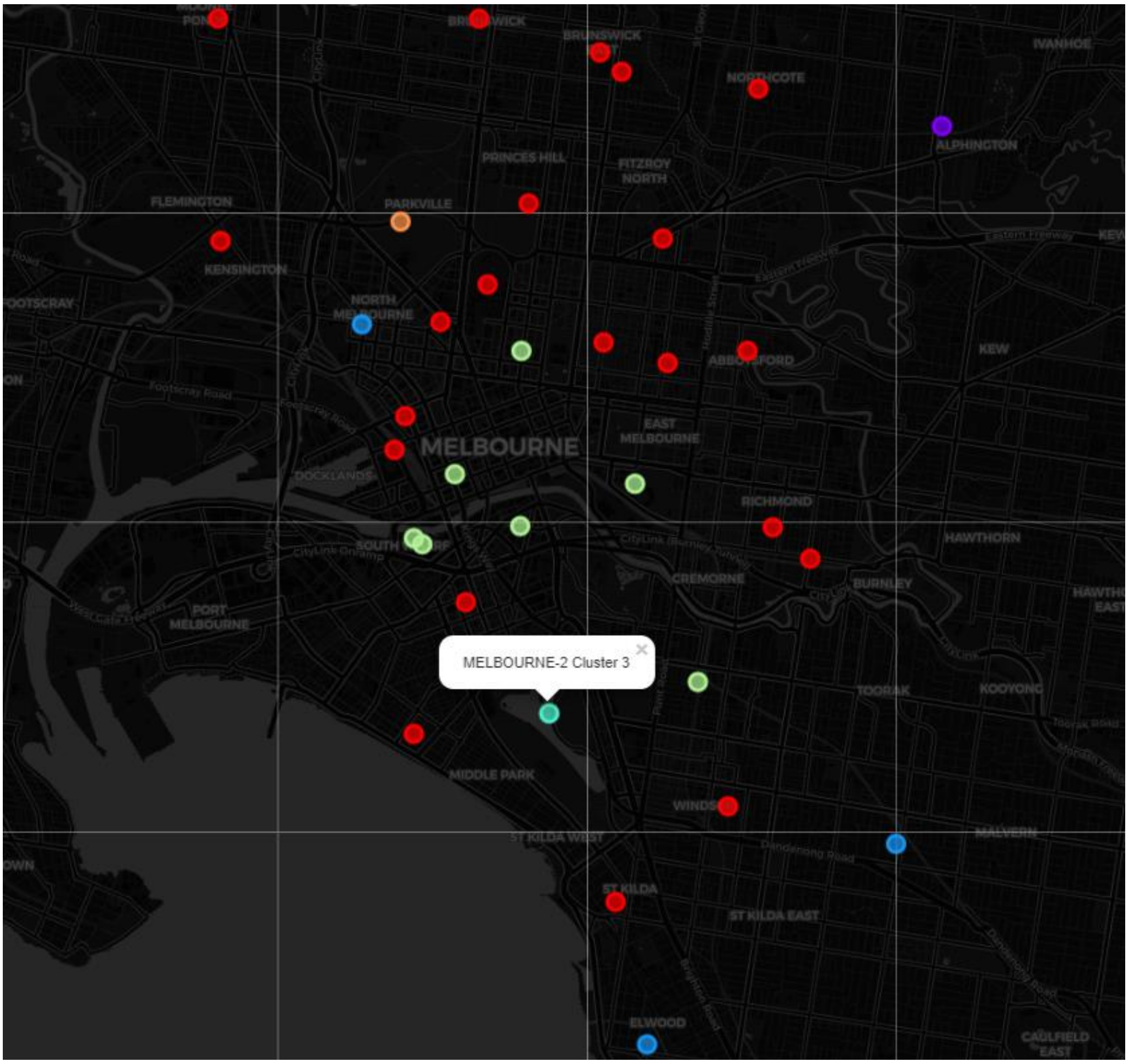


I did not get a clear result by looking at distortion. Let's look at inertia:



Unfortunately, not a significantly better outcome. In balance of the results with distortion and inertia and for simplicity, I will use **k = 6** for the purpose of this analysis. Results will be evaluated in ***Section 5. Results and Discussion***.

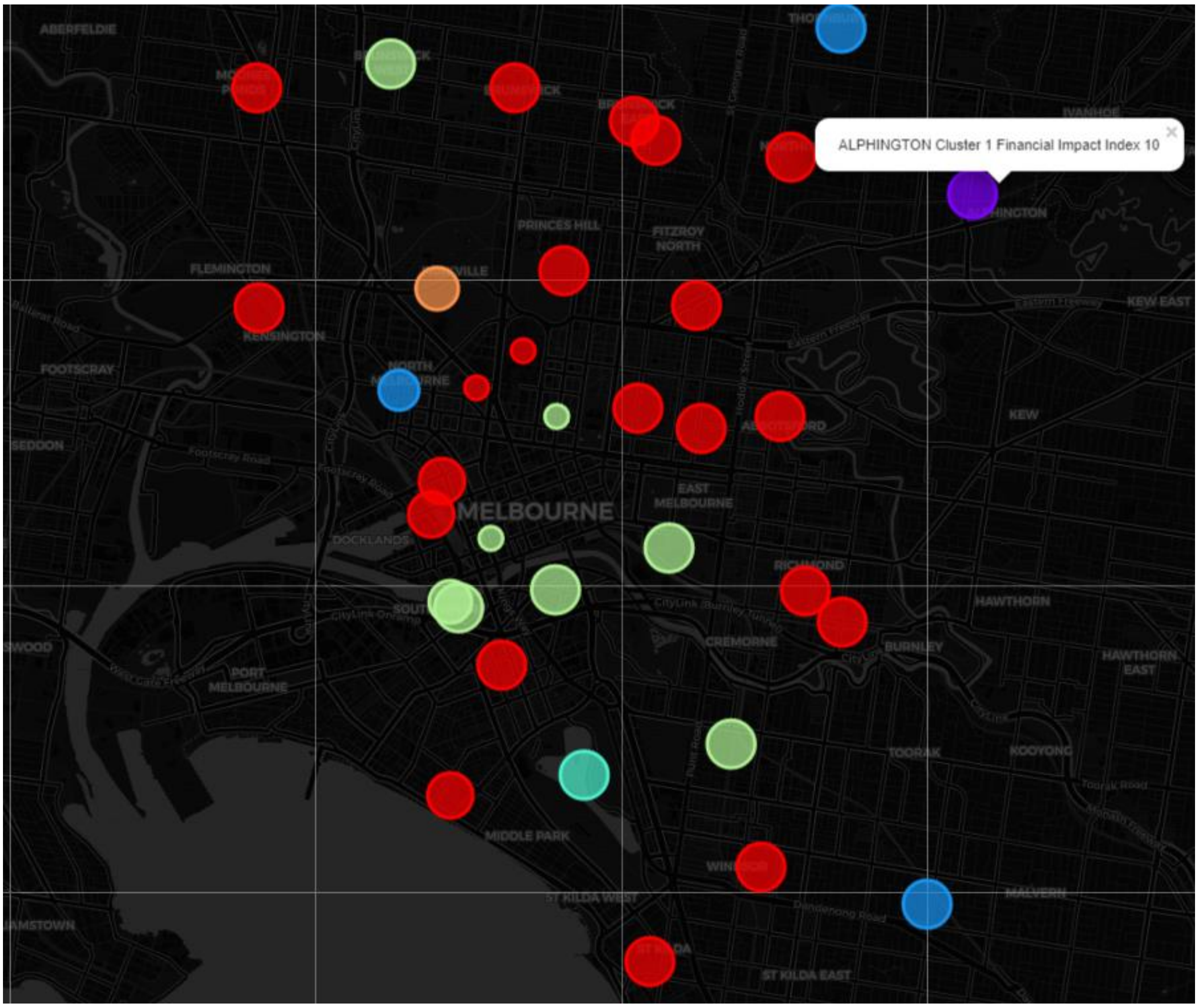
Let's now run the k-means algorithm and visualise with a map:



With results from k-means clustering, I will now merge them with the Taylor Fry COVID-19 Finance Impact Index and Median Age by suburb data sets.

COVID-19 Financial Impact index overlay

- And finally, I will visualise this on the map. The size of each 'bubble' has been scaled with the Financial Index score, i.e. the bigger the bubble, the higher the score (closer to the maximum of 10) meaning the worse the suburb has been impacted financially.



4. Analysis

- As can be seen on the map above, while the suburbs have been segregated into clusters, the financial impacts appear fairly consistent, i.e. these suburbs have mostly been severely impacted by COVID-19, with a few outliers.
- Let's have a closer look into these clusters:

Cluster	Number of Suburbs	Venue category feature
1	21	Cafés, bars, restaurants. Outliers are university and hospital.
2	1	Rental service
3	4	Stadium, stations, zoo
4	1	Breakfast spots
5	8	Clothing stores, hotels, arts venues
6	1	University, basketball court

Cluster 1

There are 21 suburbs in this cluster. Looking at the 1st to 3rd Most common venues, these are predominantly cafes, bars and restaurants. It makes sense that these have been heavily impacted by the COVID-19 restrictions. The exceptions are:
i) *University of Melbourne*: having a financial index score of 1. This suburb's main business is actually the

University of Melbourne, which was not picked up in the Foursquare data. Also noted that the median age of this suburb's population is only 20, much lower compared to other suburbs. This makes sense as university students won't be much affected from a financial perspective, plus young people with lower income have been receiving government funding support and as such are not as heavily impacted.
ii) *Royal Melbourne Hospital*: having a financial index score or 1. Again this suburb's main business is actually the Royal Melbourne Hospital, which was not picked up in the Foursquare data. Again we don't expect the hospital to have been much affected financially due to government support.

Cluster 2

There's only 1 suburb within this cluster. Most common venues are 'Rental Service' and 'Farmers Market'. These are expected to be affected financially by COVID-19, as evident from the Financial Index score of 10.

Cluster 3

There are 4 suburbs within this cluster. Other than cafes being the most common type of venues, we see a mix of other categories such as 'Football Stadium', 'Light Rail Station', 'Zoo Exhibit' and so on. Again, these venues are also heavily affected financially by COVID-19. The exception was North Melbourne faring slightly better due to having a younger population (Median age of 28) and as such may have received more government support as noted above.

Cluster 4

There's only 1 suburb within this cluster. Top venues include 'Breakfast Spot', 'Cafe' and 'Australian Restaurant'. Also in the mix are 'Zoo Exhibit and Football Stadium'.

Cluster 5

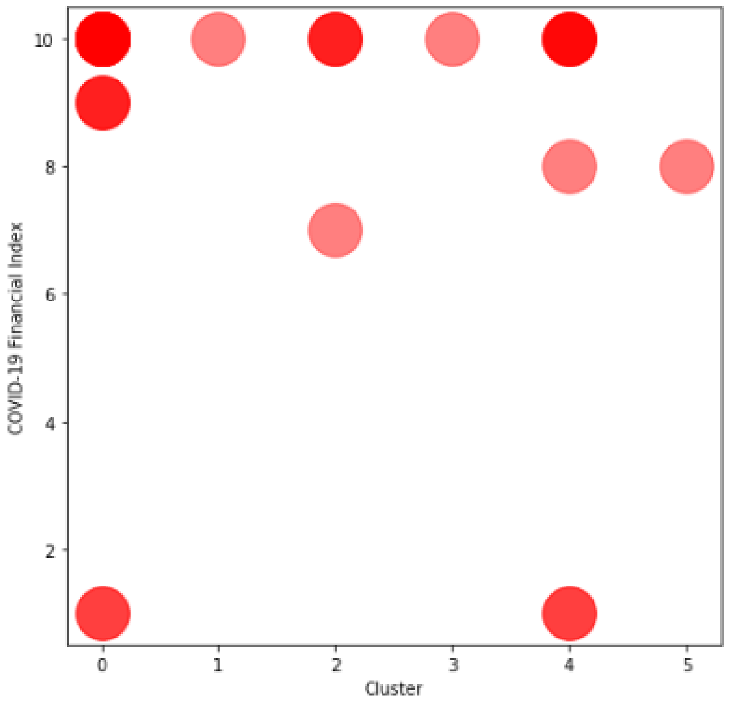
There are 8 suburbs within this cluster. This seems more of a 'shopping group', with most common venues including more 'Clothing store', 'Hotel', and 'Stores'. Again, no surprise that most of these were impacted heavily. The exceptions are Melbourne and Carlton, having a financial index of 1. As can be seen, these suburbs have a younger population and as such may have been better subsidised by the government. In addition, there's a higher concentration of offices and white-collar businesses within these central suburbs which has not been picked up in Foursquare data, which may explain why they fare better compared with others.

Cluster 6

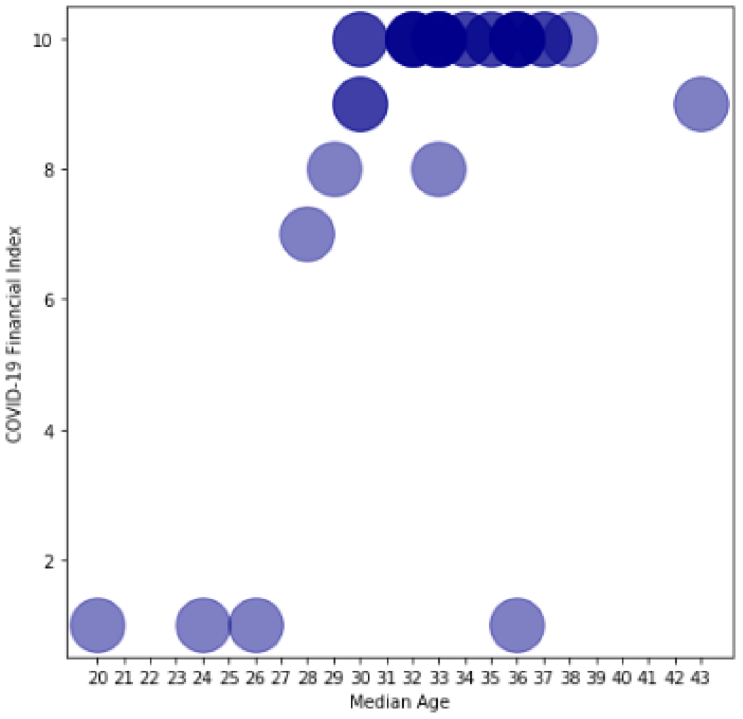
There's only 1 suburb in this cluster. Again in close proximity with Melbourne University campus, this is more of a recreational area with 'Zoo Exhibit', 'Hockey Arena' and 'Basketball Court' being featured.

Relationships

- As analysed above, we can draw some insights from the results. If we put the clusters and their corresponding COVID-19 Financial Impact Index on a scatterplot:



- We can see that most of these clusters have been similarly impacted financially by COVID-19 to a severe degree. This is consistent with my expectation set out earlier about the degree of homogeneity in these suburbs mostly having a high concentration of hospitality businesses. A few outliers have been identified as noted above, such as university and hospital being subsidised by the government and therefore seeing less of an impact, or white-collar business districts operating as usual.
- Another observation was how suburbs with a younger population were less impacted financially, possibly due to heavy government support packages to young and low income people and as such their income would not have changed substantially. Population of people above 30 years old saw a bigger impact as they would have had higher income/expenses prior to the pandemic, and disruptions/job losses would have made a more substantial change to their financials. This can be visualised in another scatterplot:



5. Results and Discussion

Our analysis shows that due to a high degree of homogeneity among Melbourne inner suburbs having hospitality businesses as the most common types of business in proximity, they have been similarly heavily affected by COVID-19 from a financial perspective. A few outliers have been identified as either having specific types of business (i.e. university, hospital, white-collar offices) or a younger population (i.e. below the age of 30) that were better subsidised by the government support packages, or were as not affected by the restrictions imposed thanks to work-from-home arrangements for example and therefore did not see a substantial change in financials.

As a result, it can be suggested that policy makers direct more support to worse affected industries such as hospitality and recreation as they have seen a bigger impact from the pandemic.

Limitations of this analysis can be attributed to the following factors:

i) *Reliability of input data*:
Only publicly available data has been used in this analysis (geographical coordinates, financial impact index, population census data). I have no assurance over the accuracy and completeness of such data. Better data means more accurate analysis.

ii) *Sample size*:
Only a sample of 36 suburbs has been used in this analysis. Expanding the sample size may draw further insights and better correlations.

iii) *Foursquare API search*:
Foursquare search has been restricted to a limit of 100 venues in proximity within a radius of 500m for simplicity within this analysis. Increasing these limits may result in more data being included and help draw further insights.

iv) *k-means clustering algorithm*:
K-means clustering was chosen for its simplicity and ease of use here. However, there are inherent limitations that I could think of:

- The variables used are of categorical nature (venue types). While one-hot encoding was used so that k-means could be applied, I expected some variance to result from the classification.
- The optimal number of cluster could not be clearly defined using the elbow method. This suggests that other clustering algorithm might have been more suitable for this analysis.
- The degree of homogeneity in sample suburbs would make it difficult to segregate them into distinctive clusters.

6. Conclusion

The purpose of this project was to demonstrate the financial impact of COVID-19 on different levels to selected suburbs in inner Melbourne, Australia. By grouping these suburbs into clusters based on a shared feature of most common types of business in proximity, a correlation could be seen that particular types of businesses such as hospitality and recreation were more heavily impacted, while others such as university and hospital were less impacted due to government subsidies. It could also be seen that younger people have been better supported by the government and as such did not see as severe of a financial impact as those above 30 years old which likely would fall into the medium-to-high income bracket and therefore saw a bigger impact.

In making the final decision, local and federal authorities should consider factors such as proximity of suburbs, specific industries and demographics, their current and future outlook to provide targeted support in balance with other public health policies in response to the COVID-19 pandemic.