Data Science Postgraduate Project
Partnership with Slalom

The Effect of COVID-19 to Melbourne Footpath Traffic

Ashwin Anis - S3763476
Stanislaus Krisna - S3703579

Table of Contents

Introduction

About Slalom

Founded in 2001, Slalom is an American business and technology consulting firm. Slalom opened their Australian office on 7 May, 2020. Slalom focuses on “strategy, technology, and business transformation”. Slalom has worked with half of the fortune 100 companies along with startups, mid-tier conganizations, and nonprofits. They employ multiple platforms to realize their projects such as Amazon Web Services and Google Cloud. They also have partnerships with Microsoft, Salesforce, and Tableau [1].

Initial Aims

The initial aim of the project was to gain insights regarding the effects of COVID-19 to Melbourne CBD footpath traffic.

Current Aims

The aim of this research is to find insights into the effect of COVID-19 towards the footpath traffic of Melbourne CBD. Then, from that insight, create a model that can predict when the footpath traffic will go back to normal. Because of some limitations, here are some of the assumptions:

- The baseline of “Normal footpath traffic” is the footpath traffic after the first lockdown (footpath traffic >= 190,000/day across Melbourne CBD). Explained on the “Footpath Traffic to COVID-19 Case Relation” results section.
- Restrictions is automatically lifted when cases reaches 0
- Everyone washes their hand and wears mask
- No new random cases. Meaning that every cases is known and counted for

Deliverables

The final requested deliverables are:

- Predictive Model
- Insights that can be gathered from the data

About the Dataset

The primary dataset that was used in this project is a public dataset that contains the number of footpath traffic across Melbourne CBD. The data is available from the public website: http://www.pedestrian.melbourne.vic.gov.au/. This data contains the following variables:

- ID: Observation ID
- Date_Time: Observation date & time (date & time)
- Year: The year the observation was conducted (number)

- Month: The Month the observation was conducted (text)
- Mdate: The numerical month the observation was conducted (number)
- Day: The day the observation was conducted (text)
- Time: The hours of observation (number)
- Sensor_ID: The unique ID of the sensor (number)
- Sensor_Name: The name of the sensor (text)
- Hourly_Counts: The amount of pedestrian recorded by the sensor (number)

The sensor location data was taken from https://data.melbourne.vic.gov.au/Transport/Pedestrian-Counting-System-Sensor-Locations/h57g-5234. The sensor location data contains the following variables:

- sensor_id: The unique ID of the sensor (number)
- sensor_description: Description on where the sensor is located (text)
- sensor_name: The name of the sensor (text)
- Installation_date: The date the sensor was installed (Date & time)
- status: Status of the sensor (Categorical, A = Active I = Inactive R = Removed)
- Note: more information regarding the sensor (text)
- direction_1: Text value of the compass first direction (text)
- direction_2: Text value of the compass second direction (text)
- latitude: numeric value, the latitude of the sensor (number)
- longitude: numeric value, the longitude of the sensor (number)
- location: Text value of both latitude and longitude (location)

The final dataset uses both sensor location data and the traffic data joined using the sensor ID from both dataset.

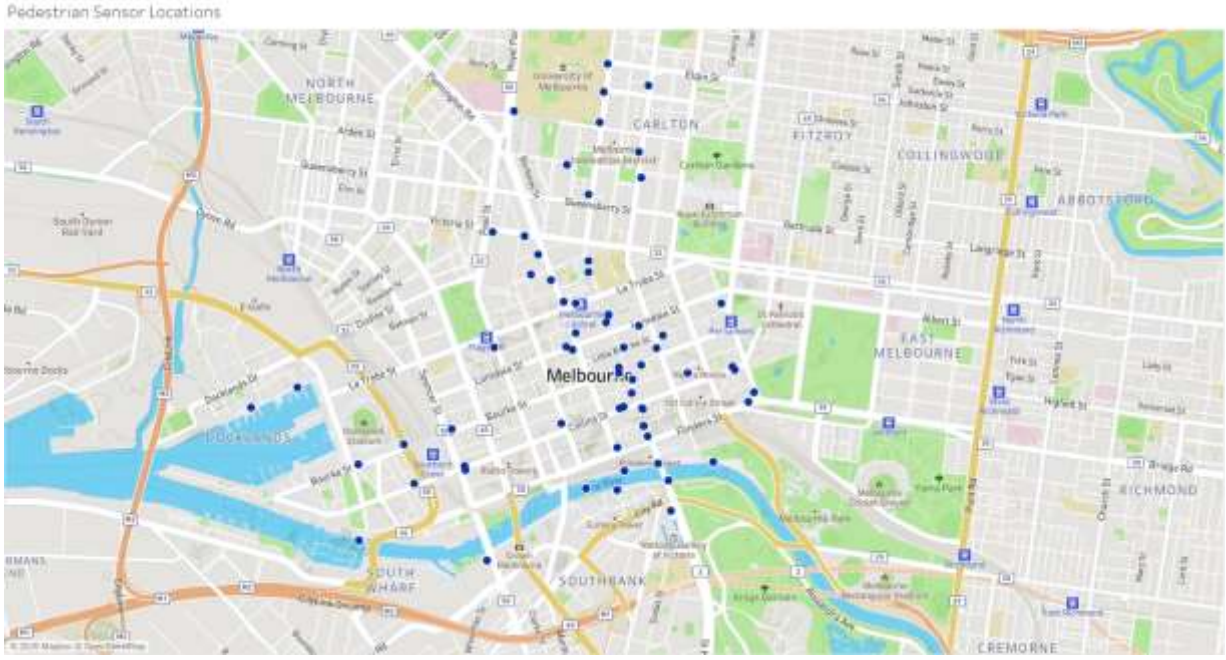


Figure 1. Footpath traffic sensor locations

The figure above shows the sensor locations all over Melbourne CBD.

Background

COVID-19 is a recent major viral outbreak that affects the whole world. The viral outbreak has disrupted business in all sectors as well as shutting down some stock markets [2, 3]. In Victoria alone, the virus has caused an estimated loss of A$2.5 billion in 2019-2020 and another estimated A$6 billion in 2020-2021 [4]. Another area that was heavily hit by the outbreak was the pedestrian traffic in the Melbourne CBD. The once bustling streets of the CBD were completely deserted due to the COVID restrictions that were in place by the Victorian Government, with data suggesting a 90 percent decrease in pedestrian traffic in August 2020 as compared to the same time last year [5].

Methodology

The methodology can be separated into 2 parts. The first part will discuss the methodology used to find insights into the effects of COVID-19 towards the footpath traffic, and the second part will touch on the methods on which the predictive model was created.

Insights

Traffic Hotspot

To find the traffic hotspot in the city, a heatmap representation of the traffic data was made and overlaid on to a map of the Melbourne CBD. To find how COVID-19 affects the traffic, the data from 2019 was scatter-plotted against the currently available data.
To plot the buildings around the Melbourne CBD area, the census data was used. The census data contains GPS coordinates for each building and the type of every building in the CBD area. After extracting the types of building and the coordinates, a scatter plot of each type of building is generated using the longitude and latitude. Then, the scatter plot was overlaid into a map to show the locations relative to the Melbourne CBD so that it would be easier to understand the visualization.

Footpath Traffic to COVID-19 Case Relation

To find the relationship, both data was plotted as a time series data. However, because of the new COVID cases and traffic volume differs, the data must be normalized. The COVID data was normalized by dividing all of the data points by the maximum number of new cases. The same process was done to normalize the footpath traffic data.

Predictive Models

While trying to find the best model to use, 2 candidate models were found. The first model that was identified was XGBoost. The other model was AdaBoost. The details of each findings will be discussed in each section. Because it is almost impossible to predict what actions the government will take, the prediction will assume that the general population will restrict themself from going outside and perform physical distancing by themself.

The variables from the data that was used to train both XGBoost and AdaBoost are:

- **New case:** New COVID-19 case recorded
- **Is weekend:** 1 if the recorded date is a weekend, 0 otherwise
- **Is weekday:** 1 if the recorded date is a weekday, 0 otherwise
- **Temperature:** temperature of the day
- **Cases from past days:** Cases from 1 day in the past until 14 days in the past. The change in new cases compared to the past days
- **Temperature past days:** Temperature from 1 day in the past until 14 days in the past). The change in temperature compared to the past days
- **Day of the week:** The day the observation falls on

Because of the relationship between the footpath traffic and the COVID-19 case (explained in the results section), to train the model, a prediction of when the COVID-19 pandemic will end (or at least under control) in Melbourne.

COVID-19 Prediction

Because this is not the main focus of this project, a single model was used to make a daily prediction. The model was created using Facebook's Prophet using the default parameter. The COVID-19 data that was used for this prediction came from the "COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University" (available here: https://github.com/CSSEGISandData/COVID-19). The daily new cases were used to train this model.

XGBoost

To train the XGBoost model, two types of data were considered. One was to shuffle the data, since the data was inversely correlated with the new daily covid data, it seems logical to treat the not as a time series data. The second type was to treat the data as a time series data and train the model using the earlier data and test using the newer datas (unshuffled).

To find the best parameter for the XGBoost model, the grid search approach was used. The grid search algorithm uses the negated root mean squared error of the model to determine the score over 10 cross-validation. The parameter to be searched are:

- seed: 0, 10, 100, 1000
- colsample_bytree: 0.3, 0.6, 1, 1.3, 1.6, 2

- learning_rate: 0.001, 0.01, 0.1, 0.5
- max_depth: 5, 10, 15, 20
- alpha: 5, 10, 15, 20
- n_estimators: 5, 10, 15, 20

The seed parameter is also searched because the XGBoost uses a pseudo random number generator. Some seeds could generate better random numbers for a certain problem. The colsample_bytree is the ratio on which a column is subsampled. The subsampling occurs every tree that is constructed. The learning rate is how much each tree contributes to the model. There is a tradeoff between learning rate and n_estimator. The n_estimator on the other hand is the amount gradient boosted trees. Max_depth is the maximum depth of each tree. Alpha is the L1 regularization term on the weights [6, 7, 8].

AdaBoost

For the same reasons as XG Boost, we decided to use two types of data – shuffled and unshuffled to train the AdaBoost model. The hyperparameter tuning was done with the help of grid search. Just like XGBoost, a 10-fold cross validation was employed.

The parameters to be searched are:

- n_estimators: 74, 75, 76, 77, 90, 120
- learning_rate: 0.001, 0.01, 0 .1, 0.5, 1.0
- loss: square , linear , exponential

Results & Analysis

Insights

There are several insights that were gained during the data exploration phase. The insights are:

Traffic Hotspot

The two figures below show the usual trends of the footpath traffic prior to the COVID-19 outbreak, specifically the year 2018 & 2019. The data consists of the average of the traffic in every hour of every day of every month. The figure below shows the average traffic on every Monday of April at 12p.m (the heatmap shows the distribution of traffic during the time of day).

Figure 2. 2018 Footpath Traffic Heatmap

Figure 3. 2019 Footpath Traffic Heatmap

When the outbreak occurs, the traffic pattern changes. According to the 2020 traffic data, the traffic is equally distributed along Swanston street and Flinders Street station.

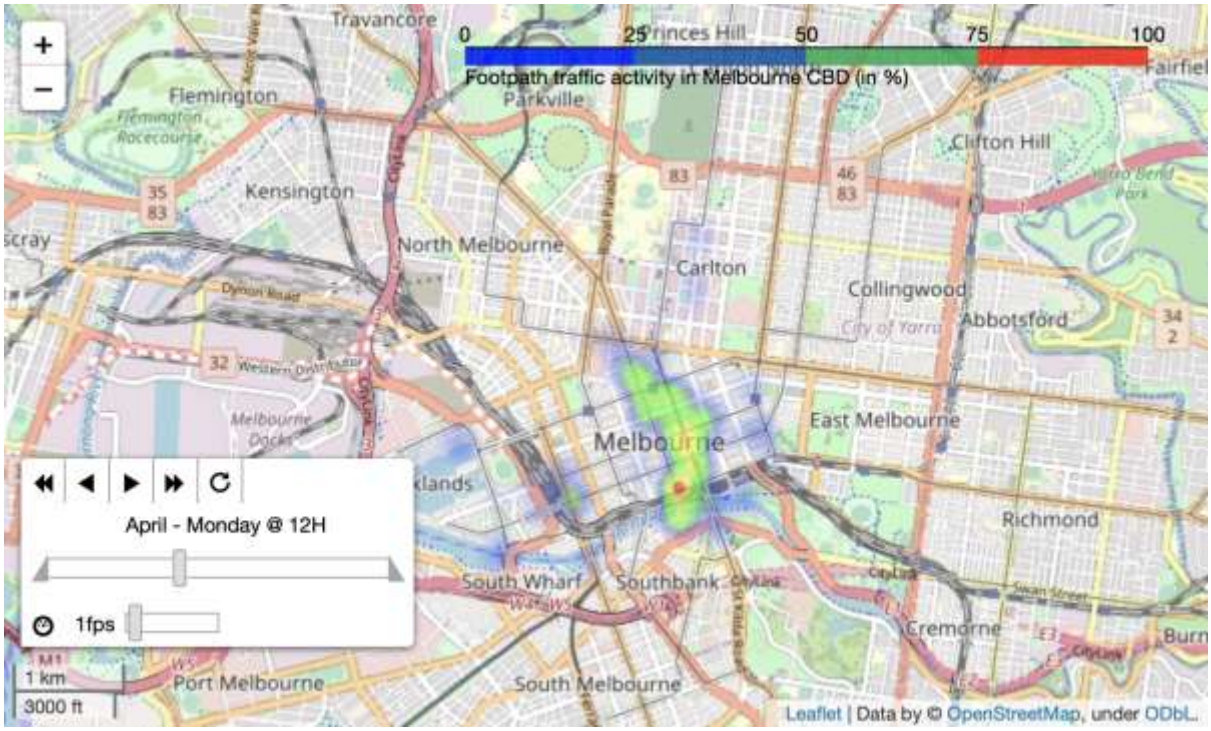


Figure 4. 2020 Footpath Traffic Heatmap

The decrease in the traffic is also backed by the scatter plot between the year 2020 and the year 2019. The chart below plots the change in traffic volume between two years. The data that is plotted are the hourly traffic data from every sensor. Because the year 2020 is still on-going, only the hourly traffic between january until July was plotted for the change in the year 2019 and 2020.

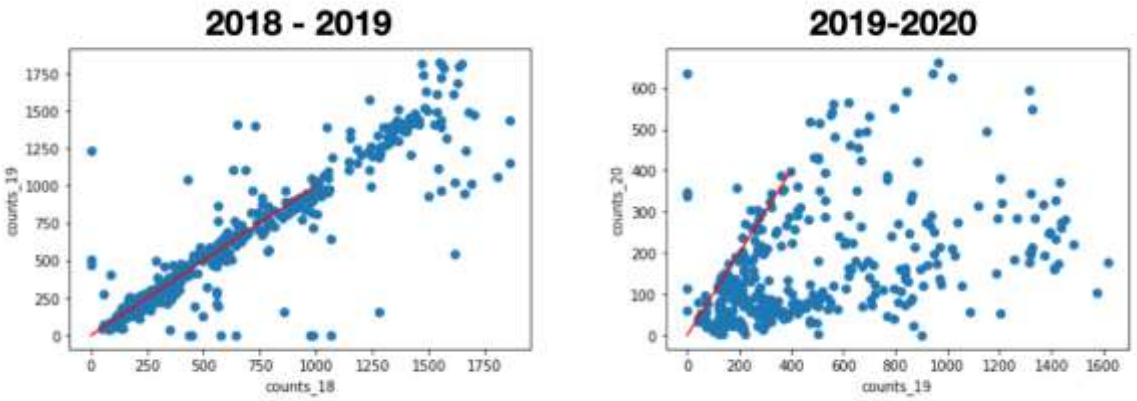


Figure 5. Change in traffic scatter plot

The scatter plot on the left shows the change in traffic between the year 2018 and 2019. On the right shows the change in traffic between 2019 and 2020. As one might expect, the change in traffic volume between 2018 and 2019 does not differ as much. On the other hand, the change in traffic between 2019 and 2020 shows that the year 2019 holds more traffic volume compared to 2020. This can be seen because most of the data points are located towards the right of the center line.

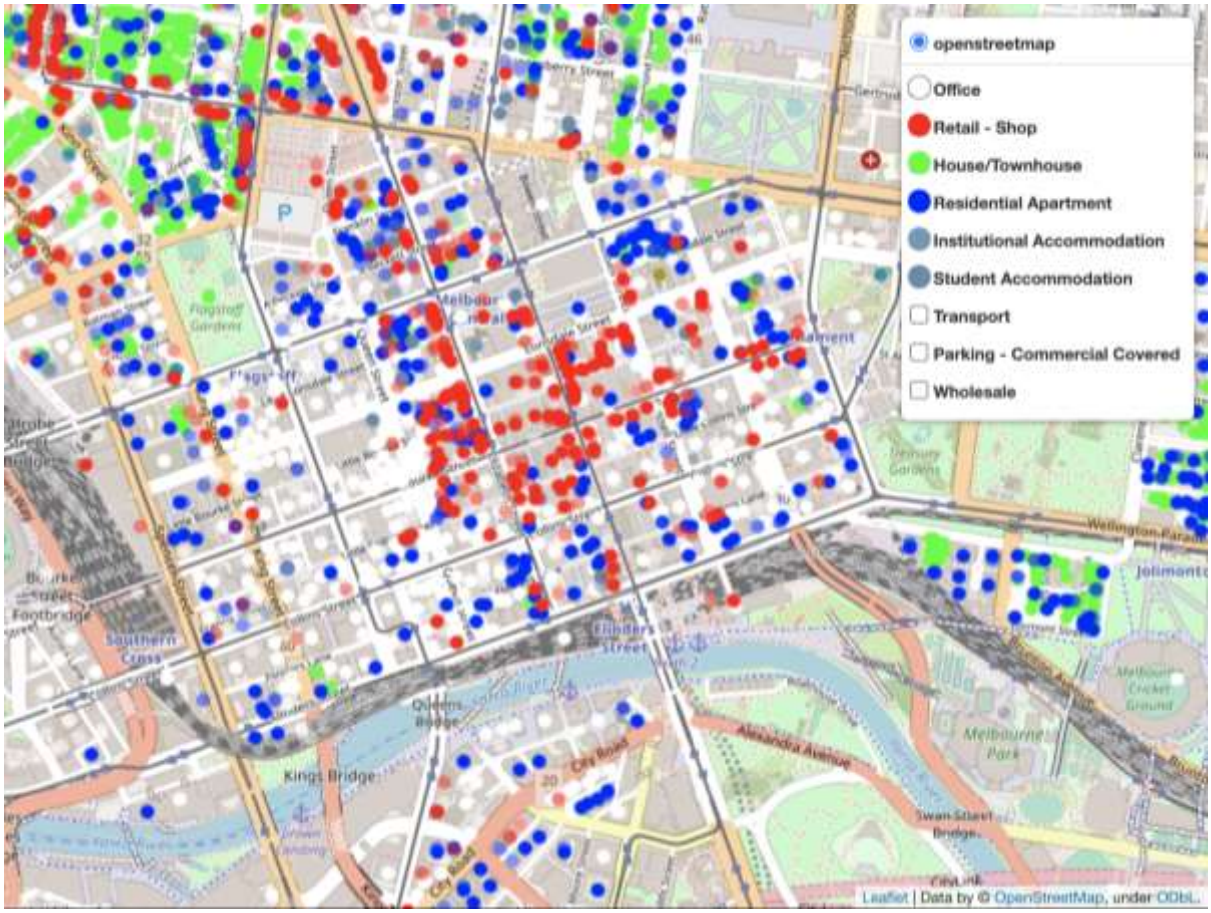
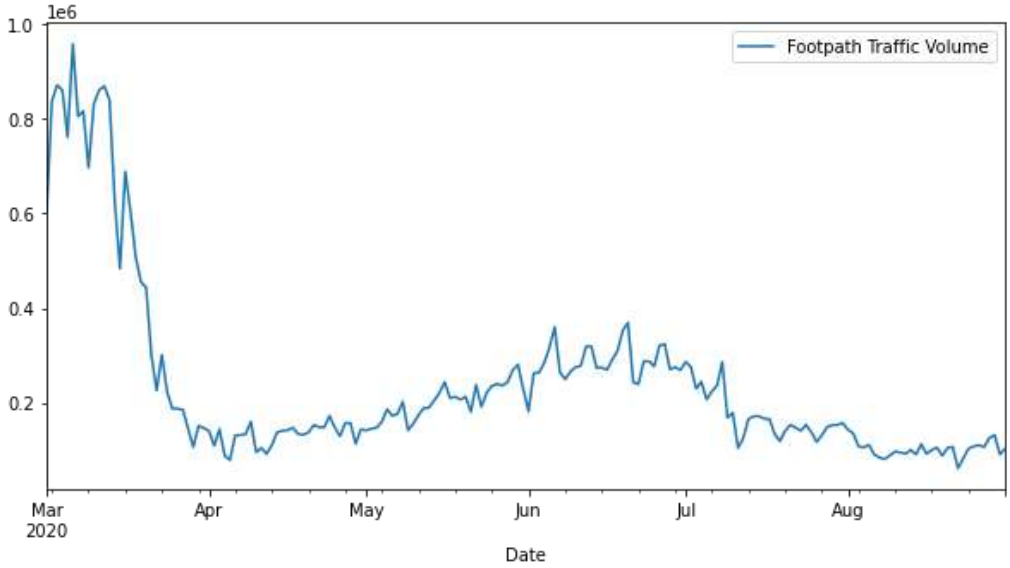


Figure 6. Types of Building Around the CBD Area

According to the table above, the types of building surrounding Swanston Street are retail shops and accommodations. Offices are excluded from the visualization because employees are encouraged to work from home. The retail shops category includes every store that sells goods or services (including restaurants, electronic shops, minimarket, and supermarket). Because most stores are allowed to open (restaurants are open for takeaway only, supermarket and minimarket are still allowed to operate, and other stores implement click-and-collect), people can still visit the shops. Secondly, because lots of apartments are also in Swanston Street, lots of people might go out to buy food, or have their food delivered to them. This explains why the traffic pattern remains constant before and during the COVID-19 pandemic. Even though there are lots of houses or townhouses in the North Melbourne area, unfortunately, since the research interest is about the Melbourne CBD area, and because the data is limited only to the CBD area, no information was available in North Melbourne.

Footpath Traffic to COVID-19 Case Relation

Figure 7. Footpath traffic volume movement (March - August).



Y axis in hundred thousands

The is an inverse relationship between the traffic and COVID-19. When the daily cases increase, the traffic count decreases and vice versa.

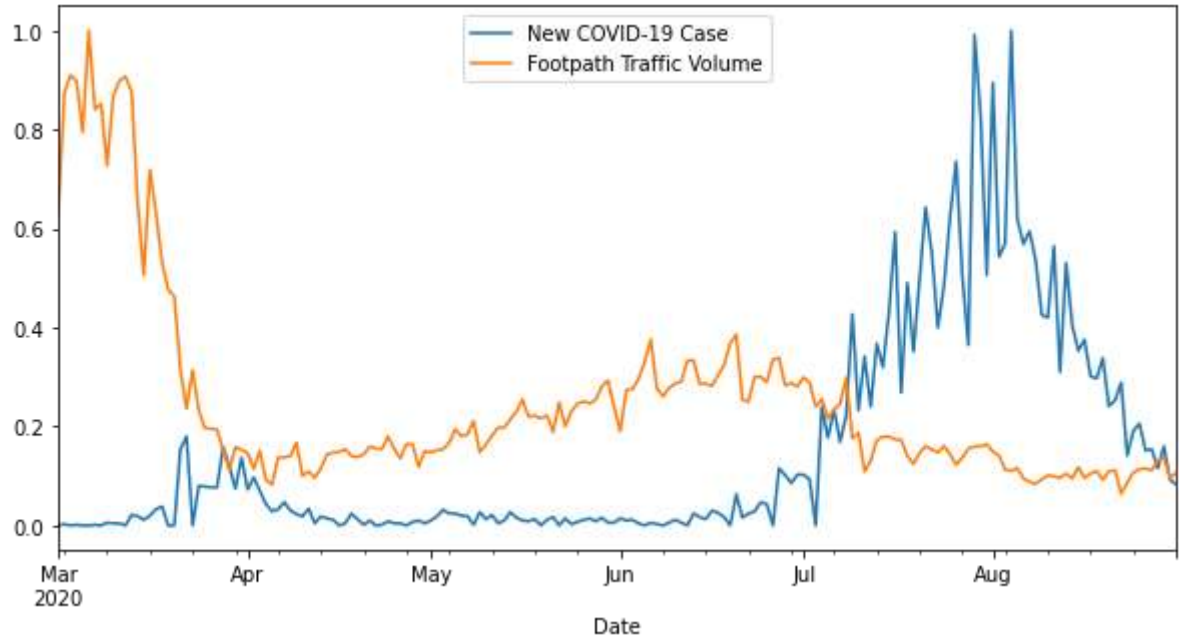


Figure 8. Footpath traffic volume to new COVID-19 cases (normalized) plot

The plot above is the normalized value of both COVID-19 new cases and footpath traffic. The normalization was performed so that it is possible to compare the relationship between the COVID-19 cases to the footpath traffic. As the table above shows, when there are more daily cases, the number of footpath traffic decreases. This behaviour can be observed between April and July (that was after the first

lockdown and before the second lockdown). When the second lockdown was introduced, the footpath traffic decreased.
190,000 footpath traffic per day is selected as the baseline since that is the value when the second lockdown starts (beginning of July) and the footpath traffic starts to decline.

Monthly Analysis of Traffic and Active COVID Cases

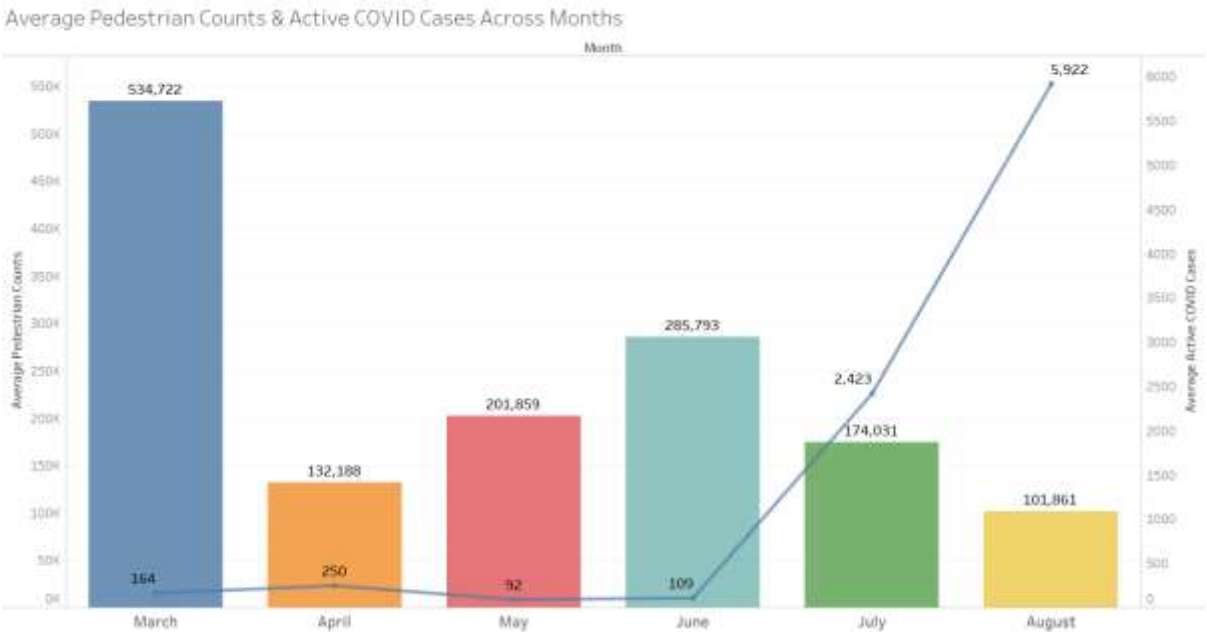


Figure 9. Average pedestrian counts & active COVID cases across months

The above graph shows the average monthly pedestrian counts from March to August. We can see that there is a huge disparity in the counts when comparing the different months. Starting with March, which marked the onset of COVID in Victoria, we can see that the average counts hovered around 534,000. This was the time when there were no movement restrictions in place and the average active COVID cases was around 164. Moving onto April, we can see that average pedestrian count has sharply declined to around 132,000 with average active cases still increasing at a steady rate. This drop in traffic is largely in part due to the state of emergency that was declared in Victoria starting from the end of March to the 2nd week of April, resulting in the restriction of public movement, banning of mass gatherings and causing the shutdown of several popular attractions in the Melbourne CBD. In May, we see that the average counts have risen to around 200,000, this also coincides with a drop in average active cases for the first time since the outbreak began. This is mainly due to the ease of restrictions that were introduced from the beginning of May as a result of the decreasing cases. This ease of restrictions resulted in people being able to freely move again thereby increasing the average counts. Moving on to June, we can see that the pedestrian counts have steadily increased to around 280,000 with a slight increase in the average active cases. By the beginning of this month most of the restrictions were removed and many businesses were starting to open their doors again.

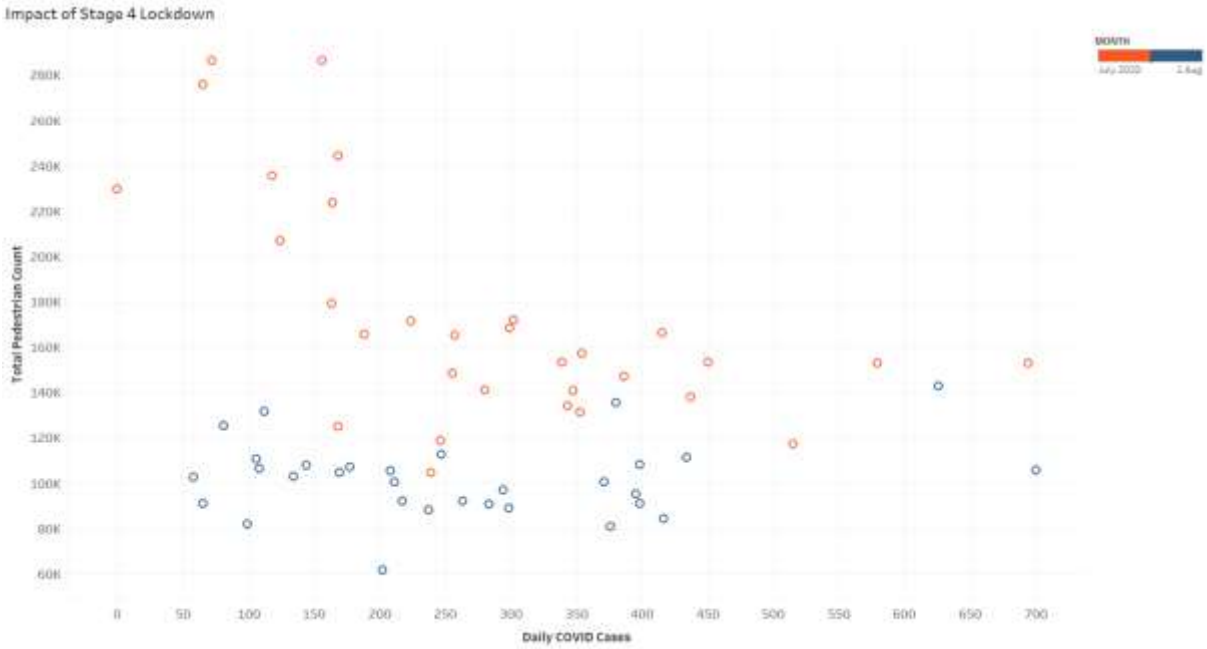


Figure 10. Impact of stage 4 lockdown

Coming to July and August, we can see that the second wave has hit Melbourne pretty badly and the city has been put under a stage 4 lockdown. The average count has decreased steadily to reach the lowest pedestrian count in August, since the start of the outbreak. This drastic change is clearly evident from the above scatter plot which shows the impact of Stage 4 lockdown on the pedestrian traffic in the CBD. The traffic for the month of July, represented by the orange dots hovers between the 140,000 to 280,000 range with an average of 174,000 while the traffic for August, represented by the blue dots hovers below the 140,000 range with an average of 101,000. This equates to a 40 percent drop in the traffic between July and August. The Stage 4 lockdown combined with a night curfew are the reasons why the pedestrian counts have shot down.

Correlation of Climatic Variables with Daily COVID Cases

As part of our search to find potential features that could be used to improve the accuracy of our model, we decided to examine the temperature and rainfall data of the Melbourne CBD to see if it has any sort of correlation with the COVID cases data. The reason we checked for correlation with the COVID data as opposed to the pedestrian traffic data is because in this scenario, the number of COVID cases is the sole reason why the city is under lockdown resulting in the low traffic. Due to this, we felt comparing the other potential features with the COVID data makes more sense. For each observation in the dataset, the temperature and rainfall reading from 6 days in the past was used. This was done since on average it takes about 5-6 days for the COVID symptoms to develop in a person, in which case they will most likely get tested around this timeframe and the case will be reported. This means that the

temperature and rainfall readings from the past might have an influence on the new cases being reported on a given day.

Temperature

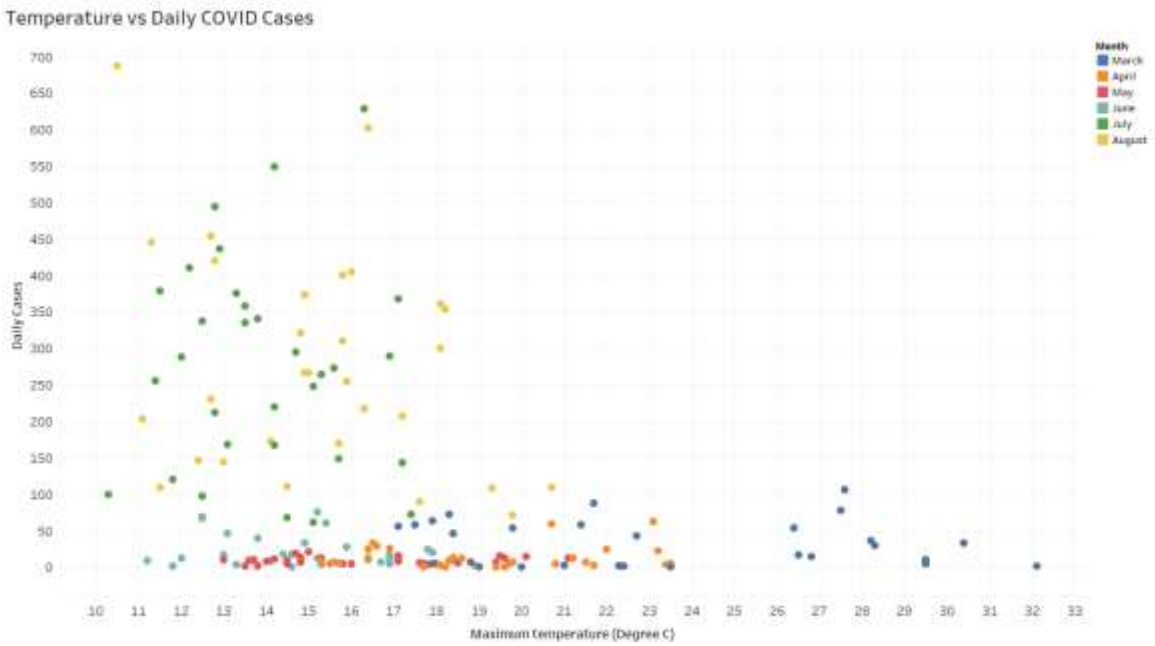


Figure 11. Temperature to daily COVID Cases

First, we examined the temperature data of the CBD. From the above scatter plot, we can see that in the months of March and April when the average temperature was about 24 degrees, the cases are on the lower end of the spectrum, while moving towards the relatively cold months of July and August when the average temperature was around 15 degrees, we can see that the cases have shot up significantly. However, this could be because of multitude of other reasons and not solely due to the change in the temperature. To test this hypothesis, we decided to conduct a Pearson correlation test to gain better understanding. From conducting the correlation test, we found out that all the correlation coefficient points to a weak negative correlation between the temperature and new COVID cases. This means, as the temperature drops, there are more COVID-19 cases. However, since the correlation is weak, this does not mean that the temperature has a direct impact towards the new COVID-19 cases.

When looking at the scatter plot, it shows that most of the data in the low temperature/cases section is grouped together. While the data points that show a high increase in the new confirmed cases are scattered. This could be explained by the recent lockdown. The end of the first lockdown was the beginning of the winter season, while the start of the second lockdown was in the beginning of July. The scattered datapoint in the high new cases area is probably the new confirmed cases

in the second lockdown. While the group in the low new cases area is probably the average new cases per day.

Rainfall

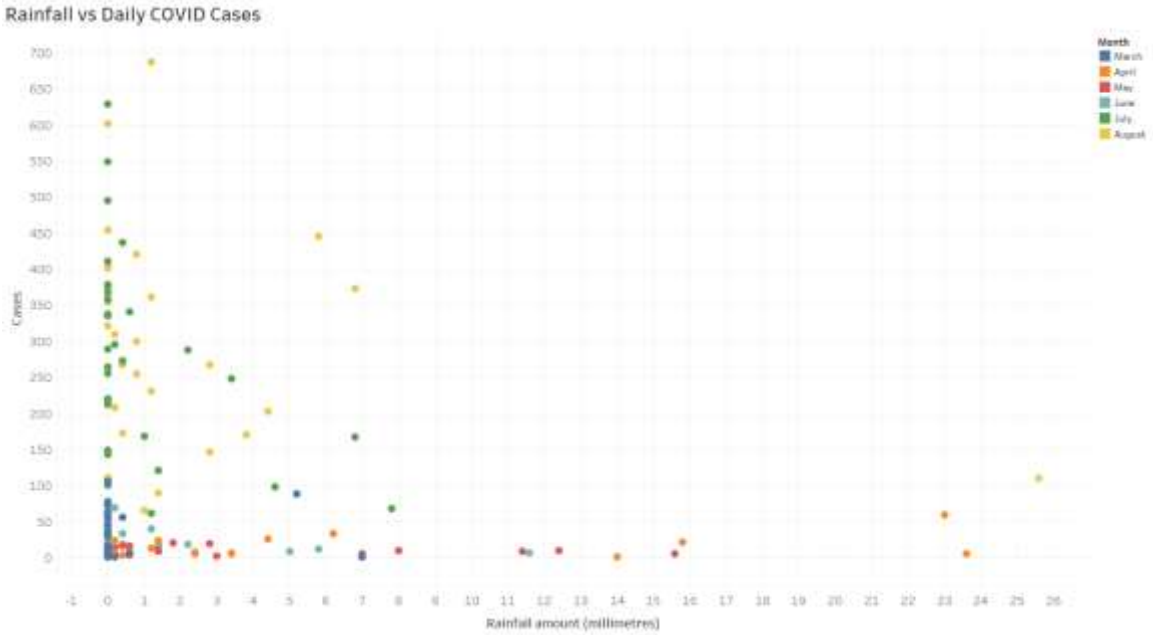


Figure 12. Rainfall Vs Daily COVID Cases

Examining the above scatter plot, we can see that the majority of the data is found towards the left end of the plot. This is a strong indication of zero or close to zero correlation between the rainfall and the COVID cases. Going through the days where there was very little rainfall, we can observe that there is a wide range of new case counts. For instance, comparing March and July, two months where there was relatively less rainfall in the CBD area. It can be seen that in March, the cases were on the lower end while for July the cases were at the higher end of the spectrum. To solidify this hypothesis that the rainfall does not have much correlation with the cases, we conducted a Pearson correlation test. The results show that the two have extremely low negative correlation between them. Due to this, we decided to not include rainfall as one of the features for the model.

COVID-19 Prediction

The model predicted that the COVID-19 daily cases will drop to 0 by the 24th of September 2020. Because the trend of the time series was taken to make the COVID-19 prediction, negative new cases will be capped to 0.

Date	**New Cases**
2020-09-17	38.232236
2020-09-18	32.254185
2020-09-19	26.276135
2020-09-20	20.298084
2020-09-21	14.320034
2020-09-22	8.341983
2020-09-23	2.363933
2020-09-24	0

This prediction does not capture the possible fluctuation of new cases that may randomly popup.

XGBoost

The results of the model training are as follows (scores shown are the mean score from the 10 cross-validation. The closer to zero, the better):

	Shuffled Data		**Unshuffled Data**	
Model #	Train	Test	Train	Test
1	-29,847.9991	-78,985.4733	-54,886.58393	-116,280.6883
2	-29,847.9991	-78,985.4733	-54,886.58393	-116,280.6883
3	-29,847.9991	-78,985.4733	-54,886.58393	-116,280.6883
4	-29,847.9991	-78,985.4733	-54,886.58393	-116,280.6883
5	-3,591.74755	-79,836.3604	-51,406.4335	-117,105.7902
6	-3,591.74755	-79,836.3603	-51,406.4335	-117,105.7902

The rows that are highlighted in green are the models that are recommended by the grid search algorithm. The parameters of the highlighted model share similar parameters. The shared parameters are:

- **Shuffled Data:**

    - colsample_bytree: 0.6
    - learning_rate: 0.5
    - max_depth: 5
    - missing: 0
    - n_estimators: 5
    - objective: reg:squarederror
    - seed: 0
- **Unshuffled Data:**
    - colsample_bytree: 0.3
    - learning_rate: 0.1
    - max_depth: 5
    - missing: 0
    - n_estimators: 20
    - objective: reg:squarederror
    - seed: 10

The only difference between each models are the alpha value, which is:

Model	**1**	**2**	**3**	**4**
Shuffled Data	10	15	5	20
Unshuffled Data	15	5	20	10

Because all of the recommended models have the same score, the only decision that has to be made is to pick the model that is trained with the shuffled data or the model trained with the unshuffled data. The difference between the train and test from the shuffled model is 49,137.4742 while the difference between the unshuffled model is 61,394.10437.

Because the shuffled model has lower average error, and because the distance between the train and test error is also lower compared to the unshuffled data, the final model is the one that is trained using the shuffled data.

Because all of the best models have different alpha values, the alpha value in the XGBoost model does not contribute enough to the model and could be disregarded and the default value of 0 could be used.

According to the model that is trained using the shuffled data, the footpath traffic would return to normal (traffic count >= 190,000/day) by the 24th of September 2020 (or on the day when the new cases reaches 0). The model assumes that the COVID-19 prediction is correct.

AdaBoost

The results from the Grid Search are as follows (R2 score is used is for scoring - closer to 1, better the score):

	Shuffled Data	**Unshuffled Data**

Model #	Train	Test	Train	Test
1	1	0.7516	1	-12.8686
2	1	0.7505	1	-12.9452
3	1	0.7502	1	-12.9562
4	1	0.7405	1	-12.9786
5	1	0.7323	1	-13.6464
6	1	0.7307	1	-15.7909

The best score from grid search is highlighted in green.Based on the r2 scores, we can see that the model does well during the training stages for both shuffled and unshuffled data. This is not too surprising to see from a decision tree-based algorithm as it tends to fit training sets perfectly. This is not the case when it comes to the testing results as there is a huge disparity between shuffled and unshuffled data. The model performs poorly on unshuffled data because the model was trained on a certain range and the test set only included a target range the model has never seen before. Hence, high negative test scores are indicative of bad performance. However for shuffled data, the model is trained across a variety of ranges (different ranges of COVID cases, temperature, traffic), this is the reason why it performs better than unshuffled data. The predictions based on the best model had a negated root mean squared error of -115,986.

The best parameter from the Grid Search is as shown below:

- **Shuffled Data:**
  - learning_rate: 0.1
  - loss: square
  - n_estimators: 77
- **Unshuffled Data:**
  - learning_rate: 0.1
  - loss: square
  - n_estimators: 77

Conclusion

There is an inverse relationship between COVID-19 and the footpath traffic at Melbourne CBD. The traffic behaviour is influenced by the type of building and the area of interest in each street. The more retail shops and apartments in a given area, the higher the traffic there is compared to areas that are mostly populated by office spaces. From conducting the correlation test for temperature and rainfall, we found

out that temperature has a weak negative correlation with COVID cases while rainfall had close to zero correlation. As a result of this, we decided to add temperature as one of the features for the model.
The best XGBoost model is the one that is trained using the shuffled data. From the unshuffled model, four models can be used. The only difference from the four models are its alpha values, which are 10, 15, 5, 20. The rest of the parameters are column sampling (colsample_bytree) of 0.6, learning rate of 0.5, max tree depth of 5, and n estimators of 5. As for the Ada boost, the model trained with shuffled data is the best model. The best parameters were learning_rate of 0.1, loss function of square and 77 as the number of estimators.
Comparing the test scores of both models, XGBoost (-78,985) had the better negated root mean squared error rate compared to that of AdaBoost (-115,986). As a result of this, the XGBoost model was chosen to make the predictions. According to the XGBoost model, the footpath traffic would return back to normal on the 24th of September 2020.

Improvement

Given more time, the predictive model could be improved. A combination of decision tree and time series analysis could improve the predictive power in both footpath traffic recovery as well as the future growth. The accuracy and model of the COVID-19 model could also be improved by investing more time on tuning and selecting the proper model. The footpath traffic sensor data is also a limiting factor. Most of the sensors are deployed on hotspots such as the train stations, tourist hotspots, and along Swanston street. A more varied sensor placement could provide more valuable insight to this project.

Appendix

Roles & responsibility

Name	Student ID	Contributions (%)
Ashwin Anis	S3763476	50
Stanislaus Krisna	S3703579	50

Ashwin Anis is responsible for the training and testing of the AdaBoost model. Stanislaus Krisna is responsible for the training and testing of the XGBoost model. Both Ashwin Anis and Stanislaus Krisna are responsible for finding supplementary data, data cleanup, and data integration. Both Ashwin Anis and Stanislaus Krisna are responsible for updating the host on project updates. They are also responsible for setting up the meetings with the host.

Self-reflection

Ashwin Anis

[Insert self-reflection]

Stanislaus Krisna

While doing this project, I learned to manage a data science project as well as learn a new decision tree algorithm (XGBoost). This project has also taught me the importance of model selection. It has also made me realize that I need to improve on my interpersonal skills. The project has broadened my horizon by...

References

1. https://www.slalom.com/newsroom/aud-25-billion-consulting-company-expands-australia
2. https://techcrunch.com/2020/03/12/covid-19-market-turmoil-tests-nyses-shutdown-circuit-breakers/
3. https://www.thejakartapost.com/news/2020/03/13/covid-19-indonesia-idx-trading-halted-first-time-2008.html
4. https://www.dtf.vic.gov.au/economic-and-financial-updates/victorian-economic-update
5. https://www.timeout.com/melbourne/news/melbourne-foot-traffic-is-down-almost-90-per-cent-since-last-year-081120
6. https://xgboost.readthedocs.io/en/latest/parameter.html
7. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html
8. https://xgboost.readthedocs.io/en/latest/python/python_api.html