# DATA*6500 Course Project Report

Evan Switzer, Namrata Roy

2024-08-16

## Introduction

Bicycle theft is a widespread problem in urban areas, and Toronto, one of Canada's largest and most lively cities, is no exception. As cycling becomes an increasingly popular mode of transportation for commuting and recreation, the frequency of bicycle thefts has risen, causing significant concern among residents, policymakers, and law enforcement agencies [1]. Toronto's dense urban layout, varied neighbourhoods, and extensive public transit system contribute to the complexity of the problem, making it challenging to implement effective preventive measures [2].

The city's dedication to promoting cycling as a sustainable and healthy transportation option is evident in its expanding network of bike lanes and cycling infrastructure [3]. However, the surge in bicycle theft poses a threat to these efforts, deterring potential cyclists and undermining public confidence in the safety and security of cycling in Toronto [4]. The financial impact of bicycle theft, coupled with the emotional distress experienced by victims, emphasizes the need for comprehensive strategies to address this issue.

Understanding the spatial and temporal patterns of bicycle theft, the socio-economic factors influencing theft rates, and the effectiveness of existing prevention measures is crucial for developing targeted interventions [2]. The goal of this paper is to gain an understanding of the factors contributing to bicycle theft in Toronto. The aim is to analyze crime patterns of bike thefts in Toronto, focusing on locations such as subway stations, bus stops, bicycle parking, etc. While our initial focus was board and included assault, auto theft, and auto theft we felt we could provide more value by focusing on bike thefts.We seek to uncover trends in bike thefts across Toronto neighbourhoods and draw valuable insights based on location from the data.

## Data Sources

We mainly used 3 datasets; Bicycle Thefts[5], Ward Profiles - Census Data[9], and Ward Profiles - Geographic Areas[9]. This data gave us all of the information we needed on bike thefts, population, and geographic areas of Toronto.

The Bicycle Thefts dataset contained detailed information on exactly where and when the bike was stolen. It also conatined extra information such as the price, make, and model of the bike, although not all of this information was filled. There are a total of 35325 records in this dataset with 35 features.

The Ward Profiles datasets contained information on the population and geographic areas of Toronto. The population dataset contained how many people lived in each ward slpit up into 5 year buckets, along with the total for each ward. Since we don't care about age we just took the total of each ward. The geographic areas dataset contained the area of each ward in square kilometers. We used this data to calculate the population density of each ward.

To supplement our data we used OpenStreetMaps to retrieve bus stops, subway stations, and bike parking stands in Toronto. This data was mainly used to analyze the proximity of bike thefts to these locations. OpenStreetMaps also let us get a layout of Toronto's wards which we used to plot the data on a map.

We also used a Major Crime Indicators [6] and Neighbourhood Crime Rates [7] at the begining before narrowing down our focus to have better analysis.

## Data Pre-Processing

The data was decently clean when we got it, but it was missing information we wanted which we got from OpenStreetMap. These were the bus stops, subway stations, and bike parking stands. We also had to calculate the population density of each ward. Most data was merged together based on the ward_id. The OCC_DATE contained a date string that was split into hour, day, month, year. There were a handful of rows with mostly NA values which were removed. The BIKE_COST column had some NA values which were replaced with the mean of the column. We considered replacing them with 0 as a missing value may mean that the owner did not value their bike.

```
tm_shape(toronto_ward_info) +
  tm_polygons(col = "Population_Density", style = "quantile", palette = "Blues", title = "Population Den
  tm_shape(BicycleThefts_sf) +
  tm_dots(col = "black", size = 0.1, alpha = 0.05) +
   tm_layout(
    legend.position = c("left", "bottom"),  # Move legend to left and bottom
    legend.outside = TRUE,  # Move legend outside of the plot area
    legend.outside.position = "right",  # Position the legend outside to the right
    legend.outside.size = 0.3  # Adjust the size of the legend (optional)
  )
```

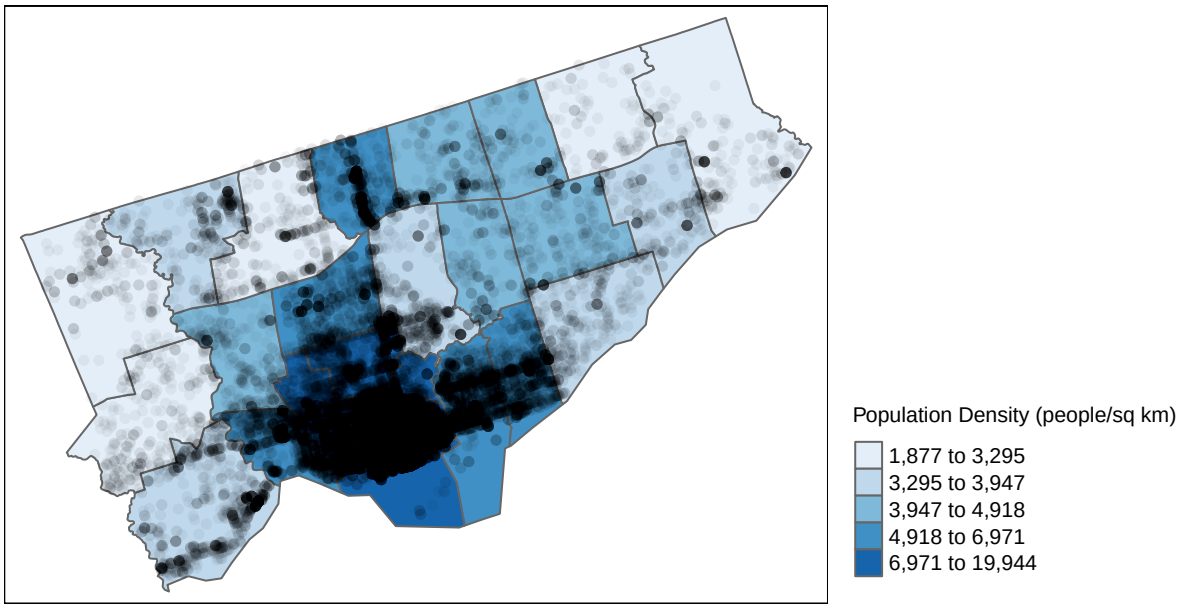



Figure 1: Population density of Toronto's wards with bicycle thefts locations overlaid.

```
#plot thefts per pop density by ward
tm_shape(toronto_ward_info) +
  tm_polygons(col = "Thefts_per_pop_density", style = "quantile", palette = "Blues", title = "Thefts per
  tm_layout(legend.position = c("left", "bottom")) +
  tm_text("ward_id", size = 1) +
   tm_layout(
    legend.position = c("left", "bottom"),  # Move legend to left and bottom
    legend.outside = TRUE,  # Move legend outside of the plot area
    legend.outside.position = "right",  # Position the legend outside to the right
    legend.outside.size = 0.3  # Adjust the size of the legend (optional)
  )
```

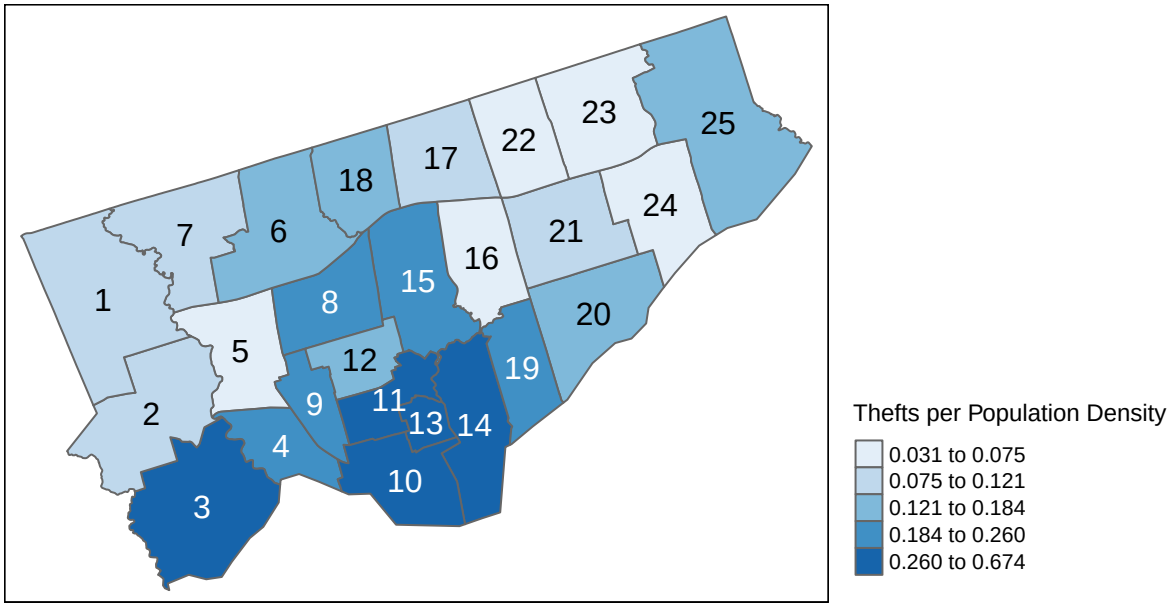



Figure 2: Thefts per Population Density by Ward.

```
tm_shape(toronto_basemap) +
  tm_rgb() +
  tm_shape(toronto_subwayStops) +
  tm_dots(col = "subwayLine", size = 0.2, legend.show = FALSE, palette = "Set1") +
  tm_shape(toronto_subwayLines) +
  tm_lines(col = "name", size = 0.5, legend.show = TRUE, palette = "Set1", col.title="Subway Line") +
   tm_layout(
    legend.position = c("left", "bottom"),  # Move legend to left and bottom
    legend.outside = TRUE,  # Move legend outside of the plot area
    legend.outside.position = "right",  # Position the legend outside to the right
    legend.outside.size = 0.3  # Adjust the size of the legend (optional)
  )
```

```
## Some legend labels were too wide. These labels have been resized to 0.61, 0.65. Increase legend.widt
```

```
#plot toronto bike parking with thefts per pop density by ward under
tm_shape(toronto_ward_info) +
  tm_polygons(col = "Thefts_per_pop_density", style = "quantile", palette = "Blues", title = "Thefts pe
  tm_shape(toronto_bikeParking) +
  tm_dots(size = 0.2, legend.show = FALSE, palette = "Set2", alpha = 0.25) +
   tm_layout(
    legend.position = c("left", "bottom"),  # Move legend to left and bottom
    legend.outside = TRUE,  # Move legend outside of the plot area
    legend.outside.position = "right",  # Position the legend outside to the right
    legend.outside.size = 0.3  # Adjust the size of the legend (optional)
  )
```

## Methods

For analysis we used Buffer Analysis, Bayesian inference(INLA), Kernel Density Estimate, and Random Forest Classification. Buffer Analysis was to visualize the proximity of bike theft incidents to subway stops,

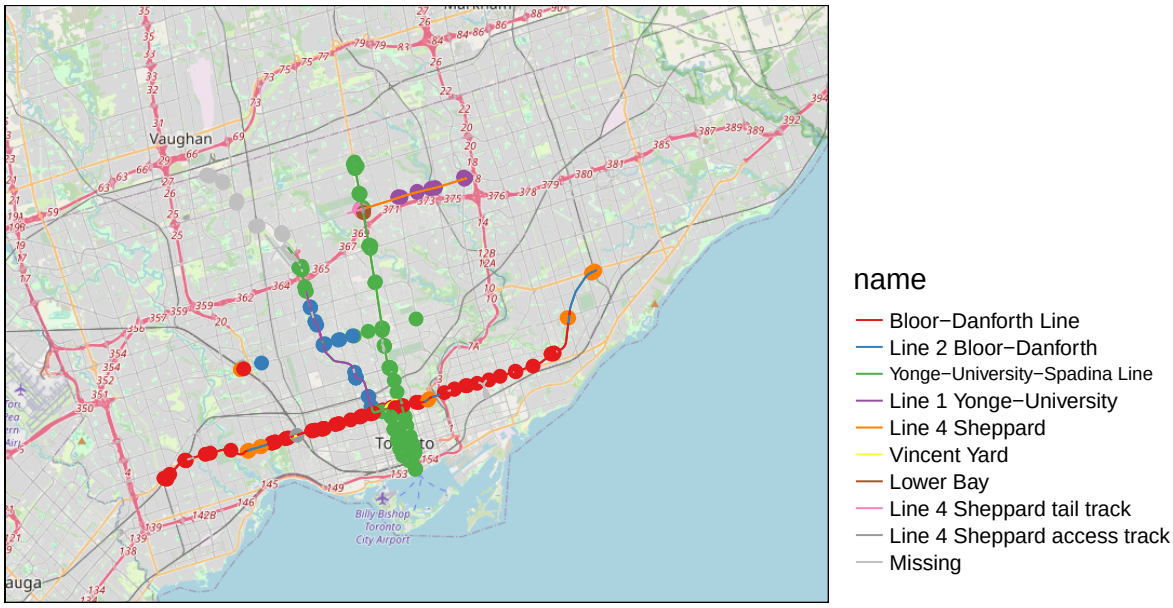



Figure 3: Subway Stops and Lines in Toronto.

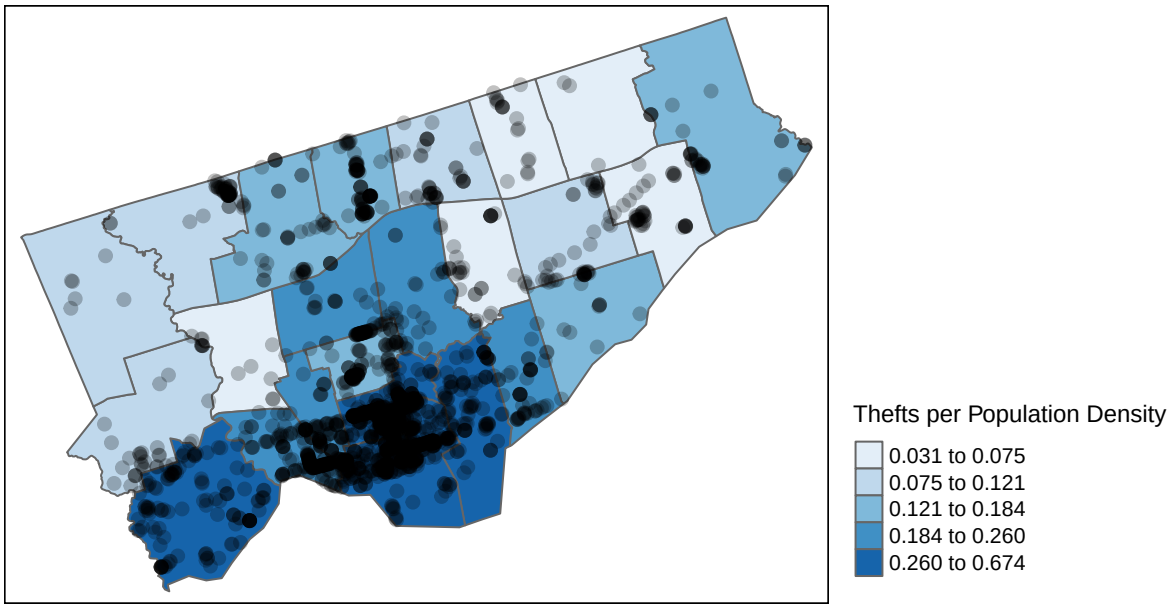



Figure 4: Bus Stops in Toronto on top of Thefts per Population Density.

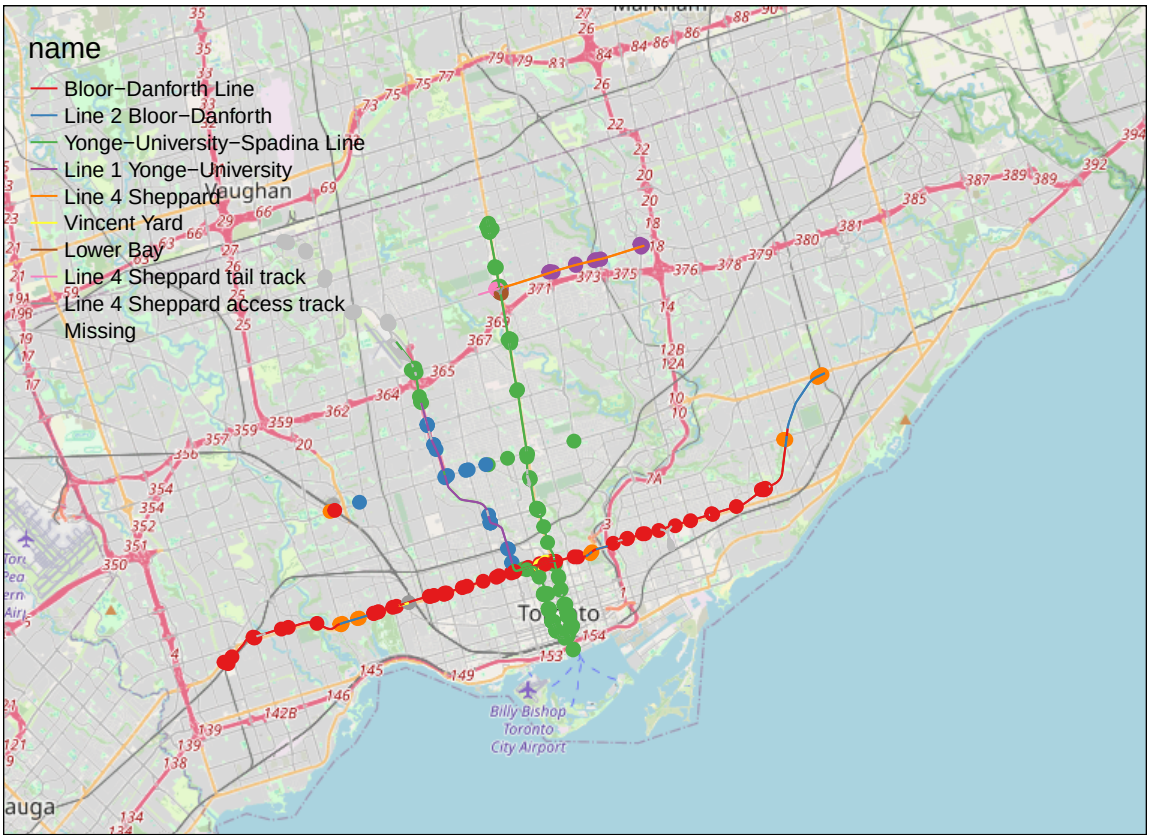



Figure 5: Background raster displaying Toronto's subway stops.

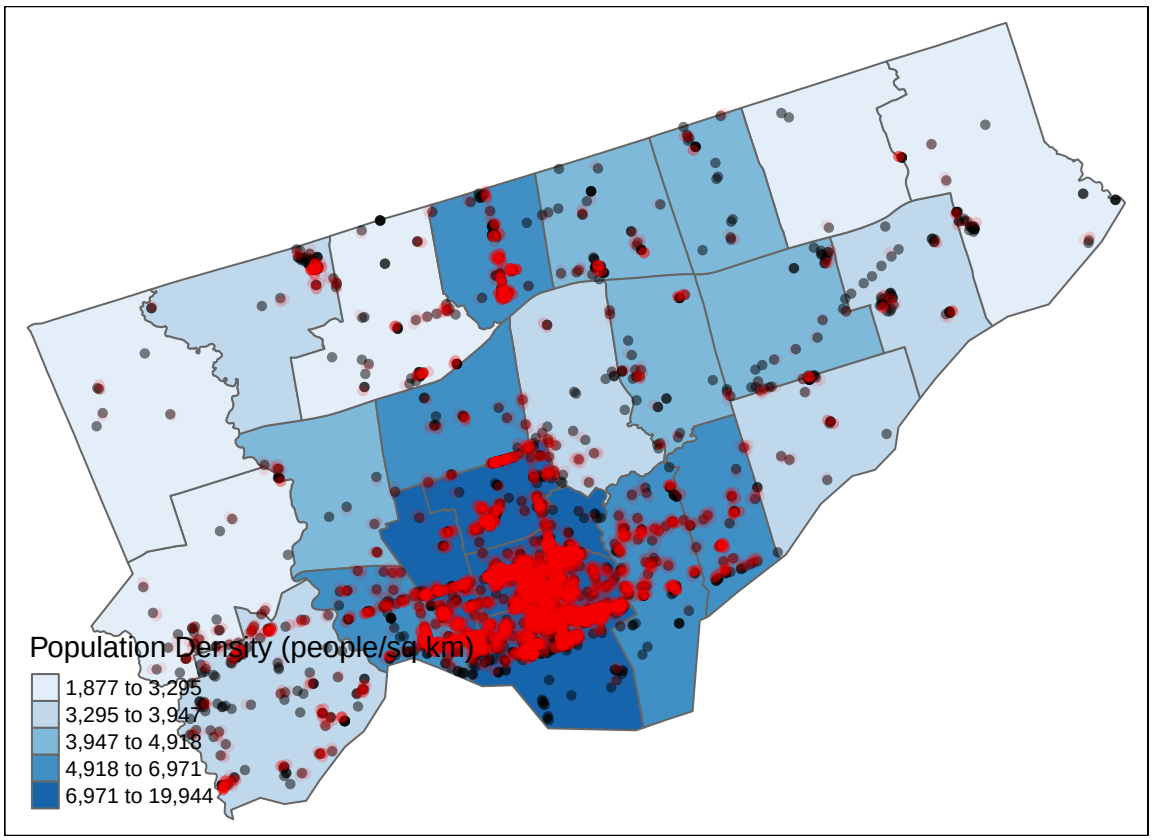
bus stops, and bike parking facilities. We used buffer analysis to create areas around our points of interest and then plot the points on a map. INLA was used to try and model the seasonal aspect of our data and see if we could predict how the data would look in the next couple years. Kernel Density Estimate was to see if there were any hot spots of bike thefts in Toronto. It was done with and with and without population density to see if there was a difference. Random Forest Classification was used to try and predict if a bike theft happened in the high density wards of Toronto or not. We grouped the highest theft wards into one group and had any thefts in that area be a 1 and any thefts outside of that area be a 0. This was our response variable.

# Results

## Buffer Analysis

```
ggplot() +
  geom_sf(data = bus_buffers, fill = "black", alpha = 1) +
  geom_sf(data = thefts_in_bus_summary, color = "red", alpha = 0.1) +
  theme_minimal() +
  labs(title = "Bike Thefts near Bus Stops")
```


```
#plot bike thefts near bike parking facilities with population density under
tm_shape(toronto_ward_info) +
  tm_polygons(col = "Population_Density", style = "quantile", palette = "Blues", title = "Population Den
  tm_shape(bike_parking_buffers) +
  tm_fill(col = "black", alpha = 0.5) +
  tm_shape(thefts_in_parking_summary) +
  tm_dots(col = "red", size = 0.1, alpha = 0.1) +
  tm_layout(legend.position = c("left", "bottom"))
```




## Bayesian Inference (INLA)

We can see that the amount of bicycle thefts spikes in the summer. This is likely because more people are outside using their bikes, so there are more bikes available to steal. There is a decrease in bike thefts in 2021, 2022 and 2023 which could possibly be explain by the COVID-19 pandemic, where people were not going outside as much. The forecast for the next 4 years does show the seasonality of the data, but does not identify a general trend of the years.

```
BicycleThefts %>%
  mutate(OCC_MONTH = month(OCC_DATE, label = TRUE)) %>%
  ggplot(aes(x = OCC_MONTH)) +
  geom_line(stat = "count", group = 1) +
  geom_point(stat = "count") +
  labs(title = "Bicycle Thefts by Month", x = "Month", y = "Number of Thefts") +
  theme_minimal()
```

```
ggplot() +
  geom_line(data = dataForcast, aes(x = YearMonth, y = forcasts), col = "red", lwd = 0.8) +
  geom_ribbon(data = dataForcast, aes(ymin = lb, ymax = ub, x = YearMonth), fill = "red", alpha = 0.4) +
  geom_line(data = dataTheft, aes(x = YearMonth, y = thefts), col = "blue", lwd = 0.8) +
  theme_bw() +
  labs(y = "Number of Thefts", x = "Year")
```

## Kernel Density Estimate

While there is the tiniest difference between the two Kernel Density Estimates, it is not enough to say that there is a difference and including population density in the KDE is not necessary. Both plots show that the highest density of bike thefts is in the downtown area of Toronto. Since we accounted for a higher population

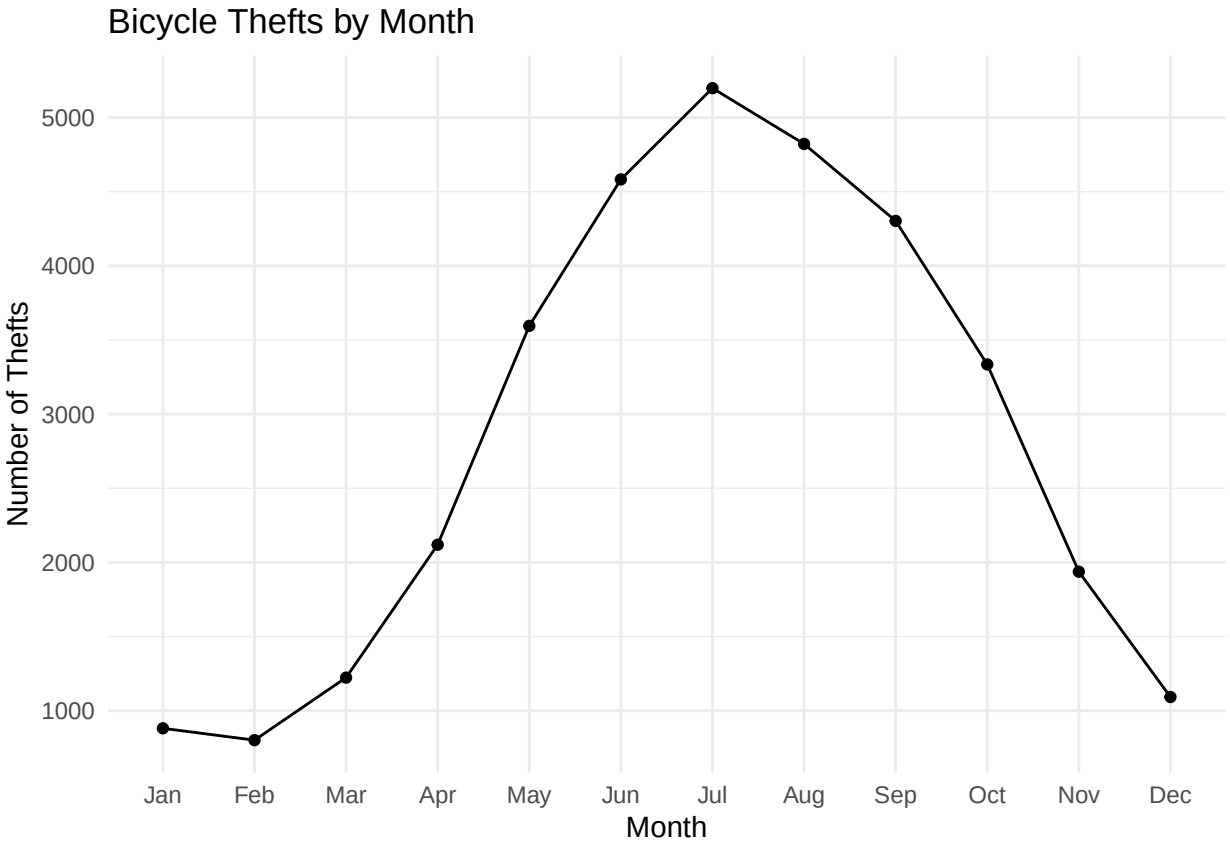



Figure 6: Bicycle Thefts by Month.

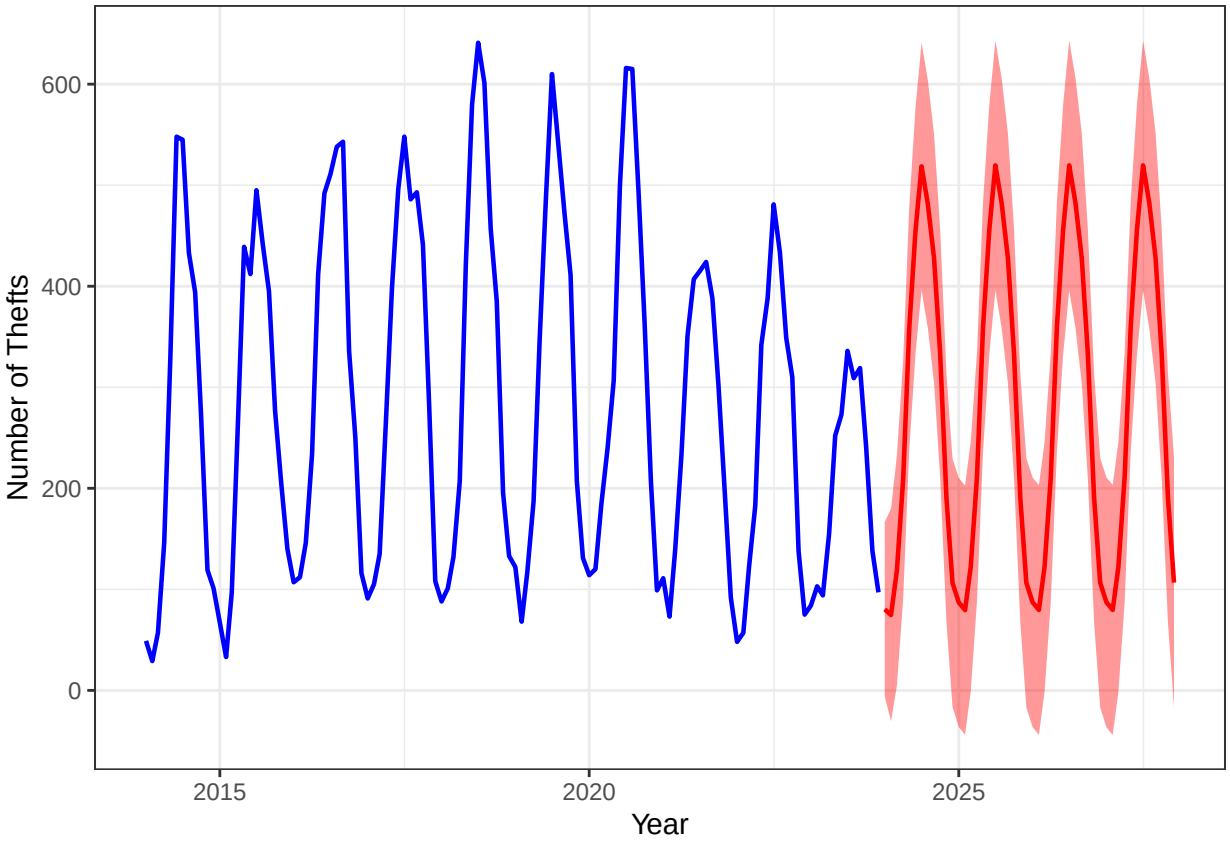



Figure 7: Bicycle Thefts Over Time with Predicted Values.

density it could be not the population of Toronto that matters, but the amount of people coming through Toronto on a daily basis

```
#plot kde
tm_shape(kde1) +
  tm_raster() +
  tm_shape(toronto_border) +
  tm_borders()
```

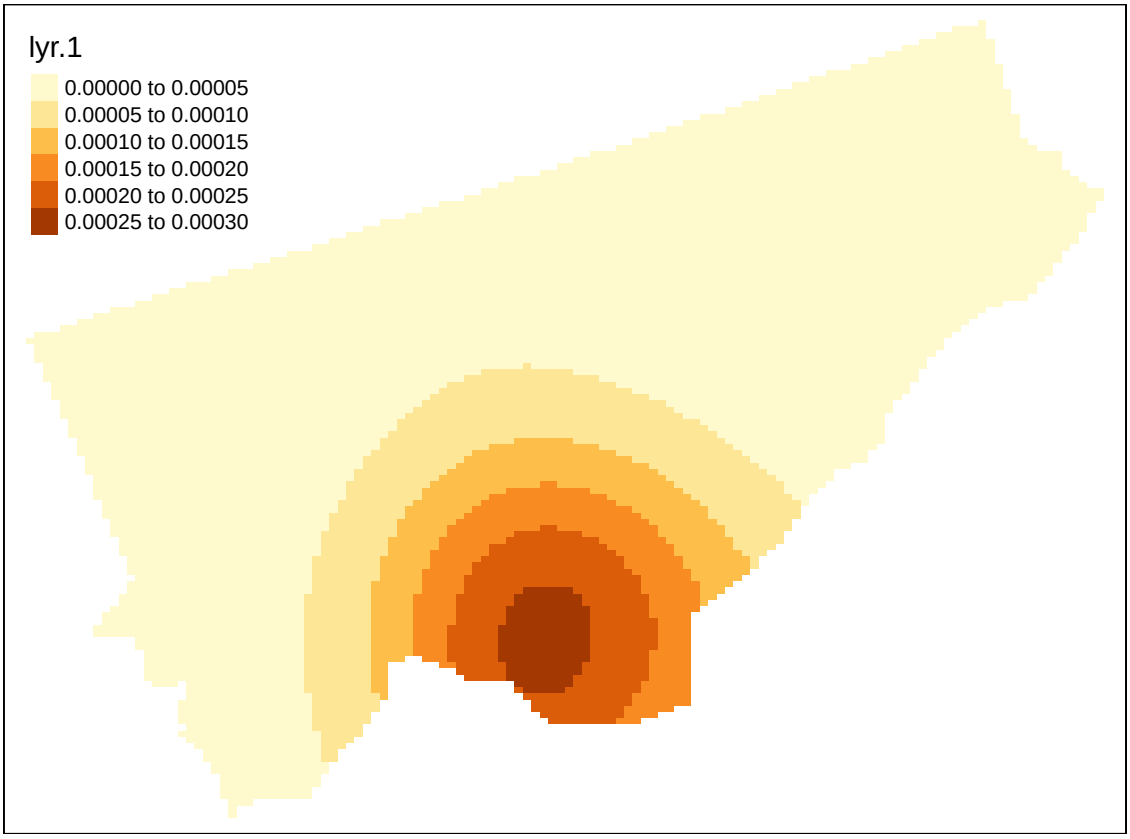



Figure 8: Kernel Density Estimate of Bicycle Thefts in Toronto.

```
# Plotting the KDE with population density
tm_shape(kde_with_pop) +
  tm_raster(palette = "YlOrRd") +
  tm_shape(toronto_border_st) +
  tm_borders()
```

```
# Convert sf object to data frame
bike_theft_model_data_df <- as.data.frame(bike_theft_model_data)

# Remove the geometry column
bike_theft_model_data_df <- bike_theft_model_data_df %>%
  dplyr::select(-geometry)

# If needed, convert it back to a standard data frame
bike_theft_model_data_clean <- as.data.frame(bike_theft_model_data_df)

#remove subways line column
```

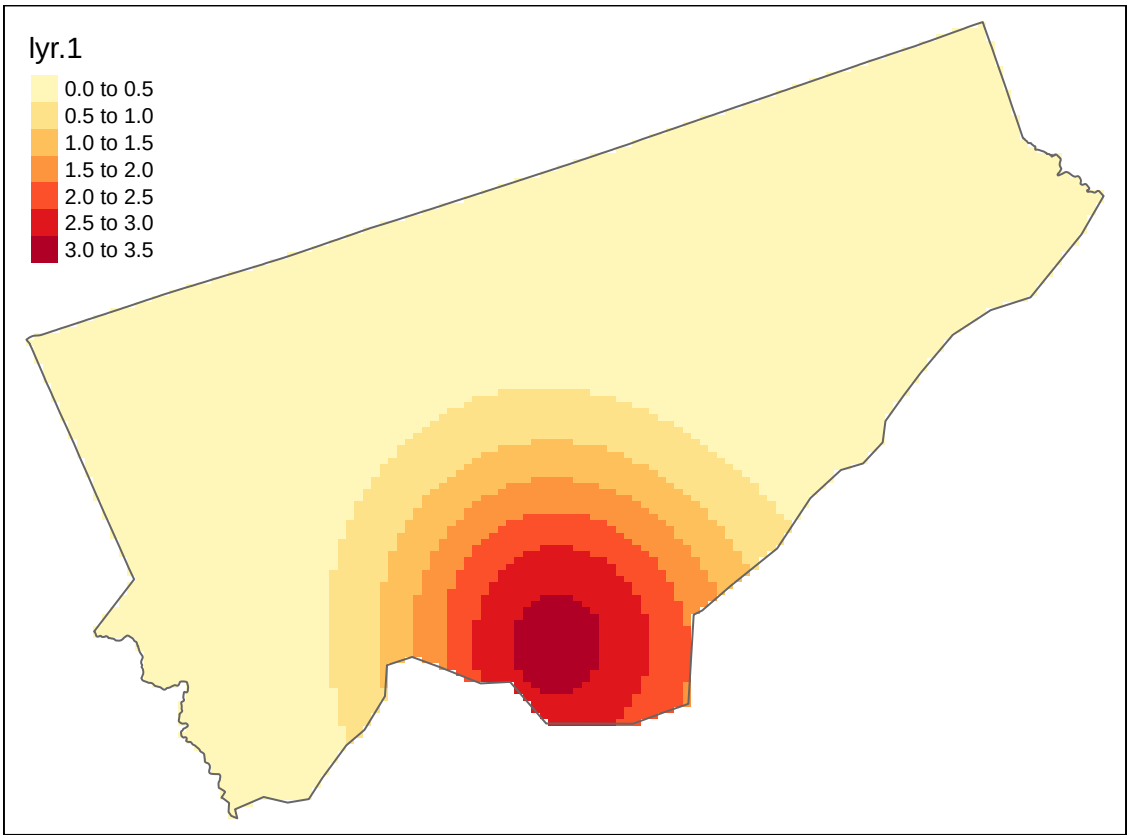



Figure 9: Kernel Density Estimate of Bicycle Thefts in Toronto with Population Density.

```r
bike_theft_model_data_clean <- bike_theft_model_data_clean |>
  dplyr::select(-c(subway_line))

#remove na values in ward_id
#bike_theft_model_data_clean <- bike_theft_model_data_clean |>
#  dplyr::filter(!is.na(ward_id))
```

## Random Forest Classification

The Random Forest Model takes into account the distance to the nearest subway stop, bus stop, and bike parking facility to predict if the theft was downtown or not. It also takes into account the type of bike, the type of premises, the day of the week, the year, the month, and the hour of the theft. The model has an accuracy of around 0.85. It shows that the most important variables are the distance to the subway and the distance to bike parking. There are more subway stations and bike parking in downtown Toronto so this makes sense. We can see the confusion matrix below that shows 5201 true positives, 792 false positives, 687 false negatives, and 3486 true negatives.

```r
# Train Random Forest model
rf_model <- randomForest(is_downtown ~ ., data = trainData, importance = TRUE)

predictions <- predict(rf_model, testData, type = "response")

confusionMatrix(predictions, testData$is_downtown)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    1    0
##          1 5196  800
##          0  692 3478
##
##                Accuracy : 0.8532
##                  95% CI : (0.8462, 0.8601)
##     No Information Rate : 0.5792
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6979
##
##  Mcnemar's Test P-Value : 0.005603
##
##             Sensitivity : 0.8825
##             Specificity : 0.8130
##          Pos Pred Value : 0.8666
##          Neg Pred Value : 0.8341
##              Prevalence : 0.5792
##          Detection Rate : 0.5111
##    Detection Prevalence : 0.5898
##       Balanced Accuracy : 0.8477
##
##        'Positive' Class : 1
##
```

```r
# Create a confusion matrix table
conf_matrix_table <- matrix(c(5201, 792, 687, 3486), nrow = 2, byrow = TRUE,
                            dimnames = list("Prediction" = c("1", "0"),
```

```
                                          "Reference" = c("1", "0")))

# Convert to a data frame for better presentation
conf_matrix_df <- as.data.frame(conf_matrix_table)

# Print the table nicely using knitr::kable
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.3.3
```

```
kable(conf_matrix_df, caption = "Confusion Matrix")
```

Table 1: Confusion Matrix

| | 1 | 0 |
|---|---|---|
| 1 | 5201 | 792 |
| 0 | 687 | 3486 |

```
print(importance(rf_model))
```

```
##                              1         0 MeanDecreaseAccuracy MeanDecreaseGini
## BIKE_TYPE             41.78084  35.80458             49.69121         402.7922
## PREMISES_TYPE         90.99813  61.36141             94.88380         587.9110
## OCC_DAY               64.73001  55.34067             82.30092        1429.5271
## OCC_YEAR              51.42613  53.95944             68.28329         601.8081
## OCC_MONTH             47.08106  44.99328             63.94877         674.3352
## OCC_HOUR              68.77067  50.13952             83.48358        1139.0856
## OCC_DOW               34.35853  36.36070             46.72328         457.4085
## dist_to_subway       117.96571  80.17481            114.05380        1862.8182
## dist_to_bike_parking 187.00474 177.05323            201.46630        3497.1148
## dist_to_bus           61.44739  65.62096             68.27754         910.1375
```

## Discussion

It seems that regardless of population density the bicycle theft rate does not increase. This goes against what we originally thought. It does seem that more foot traffic does have an effect but that foot traffic is less reflected in the population density and more reflected in the subway and bike amenities.

## Limitations and future work

Our dataset is limited by the available data on bicycle thefts in Toronto. While provided by the Toronto Police Service there is plenty of space for misreporting or under reporting thefts. A police officer who is taking the information from the victim could could have made a mistake when reporting the data. The victim could also give incorrect data, a persons memory is terrible speaking from personal experience. The dataset also had plenty of areas that were missing data or had incorrect data. For example the cost of the bike was often reported as $0. This could be because the victim valued it at $0, or did not care enough to give an accurate estimate on cost.

Another limitation we had was that the dataset only included bike thefts. Obviously it's impossible to know how many bikes were not stolen but it does mean we struggle to make models that predict bike thefts. This leads into our future work.

For future work we would like to predict bike thefts themselves. It may be possible to use population information to create synthetic data to guess what non-stolen bikes exist. Maybe there also is a dataset that

fits our needs that we did not find. We would like to create models that predict thefts using Random Forest and Gradient Boosting. We would also like to use the data to create a risk map that shows areas with a higher predicted risk of bike theft. This could be useful for police to allocate resources to areas that need it most.

# References


[1] Toronto Police Service. (2020). Bicycle Theft Statistics Report.

[2] Piza, E. L., & Kennedy, L. W. (2012). The impact of sample size on the performance of hot spot identification techniques. Applied Geography, 34, 255-270.

[3] City of Toronto. (2021). Toronto Cycling Plan: 2021-2025.

[4] Nettleton, S., & Green, J. (2014). Urban cycling and the politics of mobility: Critical perspectives on policies and practices. Transport Policy, 35, 232-239.

[5] Major crime Indicators open data. (n.d.). https://data.torontopolice.on.ca/datasets/0a239a5563a344a3bbf8452504ed8d68_0/explore

[6] Neighbourhood crime Rates open data. (n.d.). https://data.torontopolice.on.ca/datasets/ea0cfecdb1de416884e6b0bf08a9e195_0/explore

[7] Toronto Police Service Public Safety Data Portal. (n.d.). https://data.torontopolice.on.ca/maps/a89d10d5e28444ceb0c8d1d4c0ee39cc

[8] OpenStreetMap contributors. (Year). OpenStreetMap. Retrieved from https://www.openstreetmap.org

[9] Open Data Dataset. (n.d.). City of Toronto Open Data Portal. https://open.toronto.ca/dataset/ward-profiles-25-ward-model/