Long Mai
Josh Vocal
Kevin Antonius

Vancouver Bike Theft

Computational Data Science

CMPT 353

Fall 2019

Simon Fraser University

ABSTRACT

Biking is one of the cleanest and most energy-efficient forms of transportation and the number of people choosing to cycle continues to grow every year. Bike thief is a common problem that everyone who owns a bike needs to be aware of. In Vancouver, bike parts are often stolen and are resold for high prices. This project aims to help reduce the number of bicycle thefts in the City of Vancouver by examining external factors that may be correlated to the number of bicycle thefts such as time, weather, and location.

DATA

There are two datasets that we used: Vancouver Crimes from VPD Open Data and Vancouver Weather from the Government of Canada. Since both datasets are from the government, it can be assumed that our dataset will be reliable for our analysis. The Vancouver Crimes dataset contains a list of different types of crimes with the time and location that they occurred. One type of crime we are particularly interested in is the type Theft of Bicycle. The Government of Canada weather site contains 2 different types for the dataset: weather monitored hourly and weekly. Since only the hourly dataset contains the weather status for each hour, which is what we needed, we obtained it via the site's API to incorporate it into our analysis.

CLEANING

For the Vancouver Crimes dataset, we filtered out the other types of crime and only included Bicycle Theft. The dataset included UTM coordinates for each of the crimes which were more difficult to apply than latitude and longitude. We converted these coordinates into latitude and longitude. There were separated columns for year, month and day. We added a new column which would be a Datetime object to allow us to plot time more easily. Finally, we dropped any rows which included NA values and sorted the entire dataset by year.

For the Vancouver Weather hourly dataset, we only selected the date, time and weather status as they are needed for us to determine the weather status for each theft occurred. However, there exist a lot of entries with missing values for the weather condition. Since we cannot make an assumption about the type of weather for those hours as it would make our data inaccurate and bias, we decided to drop those entries with missing values. We also added a column to our weather data that indicates whether the condition of the weather is good or bad. By using the dataset's documentation, we define good weather conditions as Mainly Clear, Mostly Cloudy, Cloudy, Clear, since these type of condition does not hinder the user's ability to ride their bike.

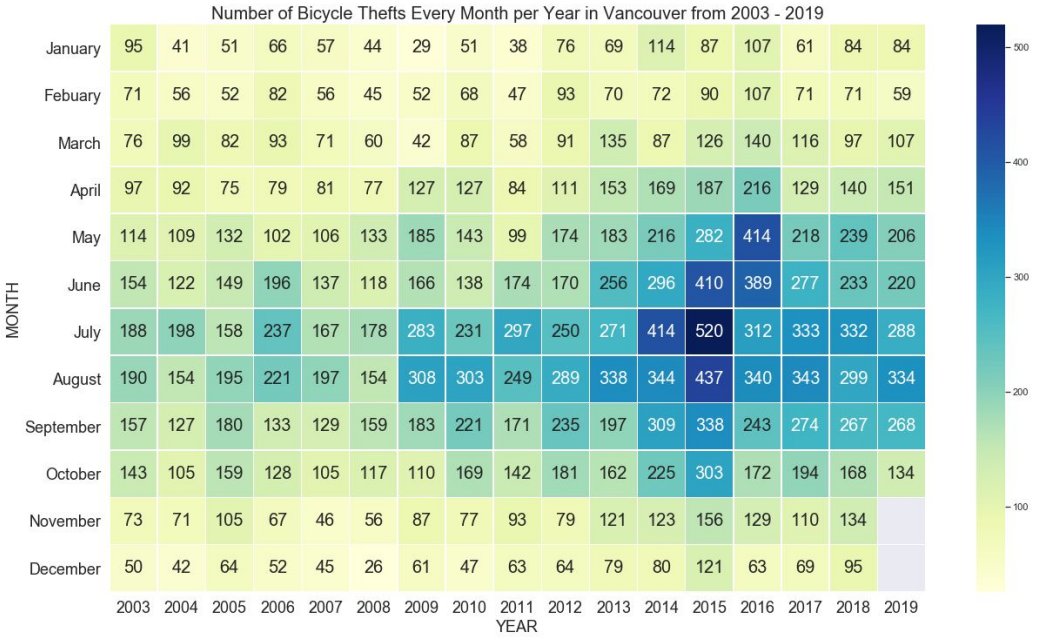
TECHNIQUES

In our exploratory analysis, we used various methods to understand the data. Using linear regression, we were able to understand the relationship between the number of thefts over time. The normality test and t-test were used to understand if the means were different between the year ranges as well as the weather conditions for the bike theft. In order to present the data, visualization libraries such as matplotlib and seaborn were used to create the charts and graphs.

In our predictive analysis, we trained various machine learning models to try and predict the hour theft will occur. Decision Tree, Random Forest, Naive Bayes and Gradient Boosting Classifiers were used. In order to optimize the accuracy of our models, a library called GridSearchCV was used to find the best parameters for each model.

RESULTS AND VISUALIZATIONS

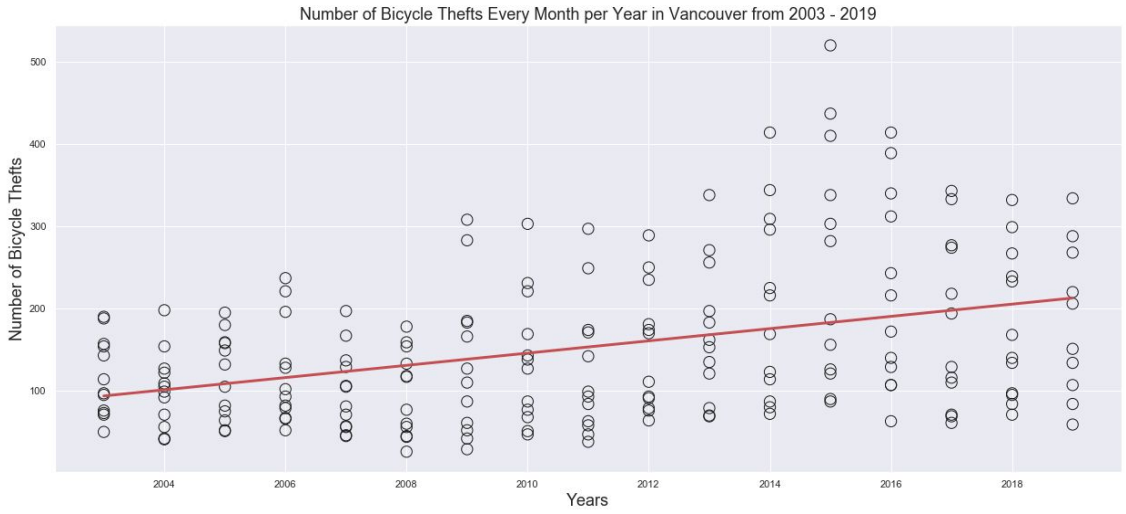
1. Exploratory Analysis



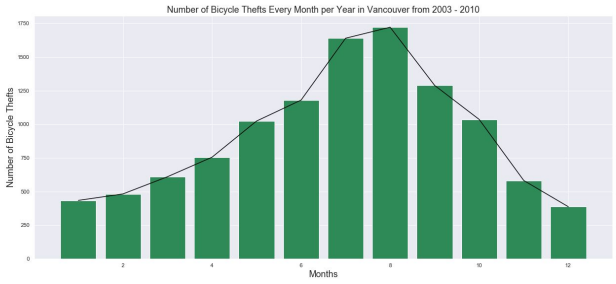
Here's an overall view of the number of bike thefts each month per year in Vancouver. By looking at the heatmap, we can determine a few things. We can tell that 2015 was one of the worst years for theft. The months of June, July, August, and September are where most thefts occur. The trend of bike thefts looks like it is increasing over the years. We will take a closer look at all of these observations to determine if they are true.

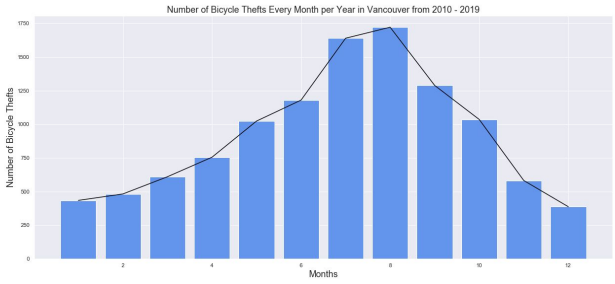
Is Bicycle Theft Increasing in Vancouver?

Our hypothesis is that bike theft is increasing over time in Vancouver. We want to determine the relationship between the number of bike thefts occurring every year. By plotting the number of bikes stolen every month per year, we created a scatter-plot to analyze the data. Using linear regression, we can model the relationship between the number of bike thefts over the years. By using the slope and intercept from our calculated linear regression, we are able to plot the best-fit-line. Taking a closer look at the best-fit-line, we can see that it is gradually increasing over time.



Another value we can analyze in our linear regression is the p-value. A low p-value, which is $p < 0.05$, allows up to reject the null hypothesis. On the bottom-right corner of the graph, the p-value is 1.79e-08 which well below 0.05 so we can safely reject the null hypothesis. We can determine that bike theft is increasing over time in Vancouver.

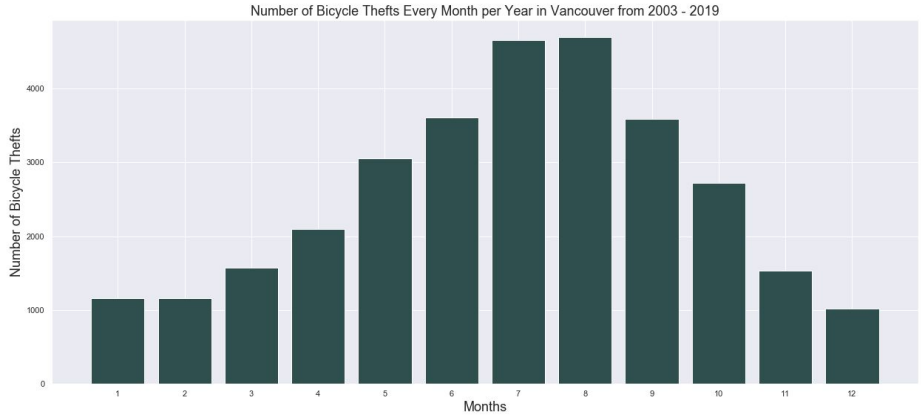




We want to determine if there is a difference in the average number of thefts between the previous decade. The green plot is the total number of bike thefts from 2003 - 2010. The blue plot is the total number of bike thefts from 2010 - 2019.

One statistical test we can perform is the t-test to determine if the means are different. First, we need to determine if both graphs are normal to perform the test. It turns out that both of these are normal with the green graph having a p-value of 0.50 and the blue graph a p-value of 0.27. We have a $p > 0.05$ so we don't have to reject the test. Using the t-test, we determined that the p-value is 0.021.

What Month Does Theft Most Often Occur?



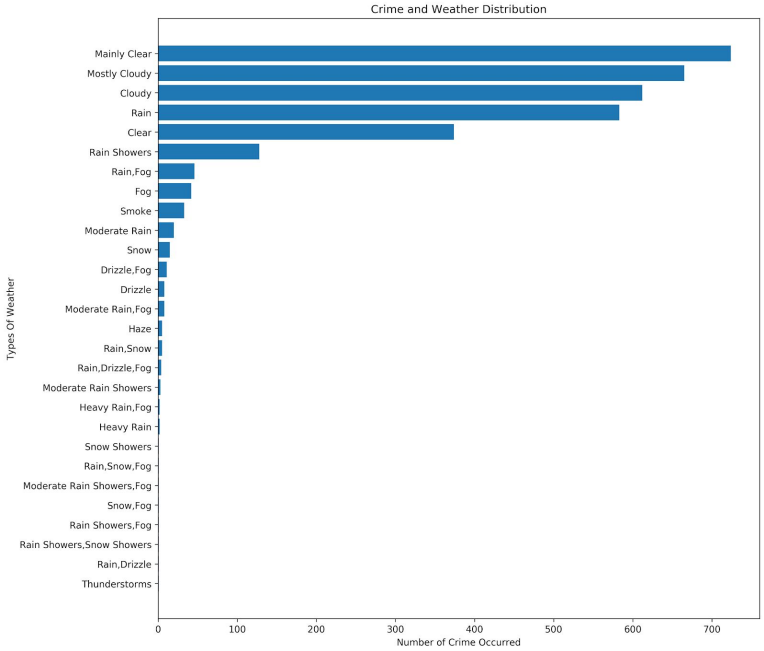
Normal Test P-value: 0.3593806206983924

We believe that a majority of the bikes are stolen during the summer in Vancouver because that's when people are most likely to ride their bikes. By plotting out the total number of thefts each month, we can see a peak. The Fall and Spring seasons are typically rainy, there will be fewer bikes to be stolen since not many people are riding their bikes. The months of June, July, August and September have historically the most number of thefts.

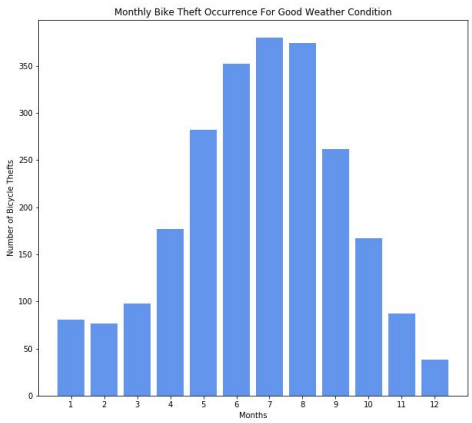
What Time During The Day Does Theft Most Often Occur?

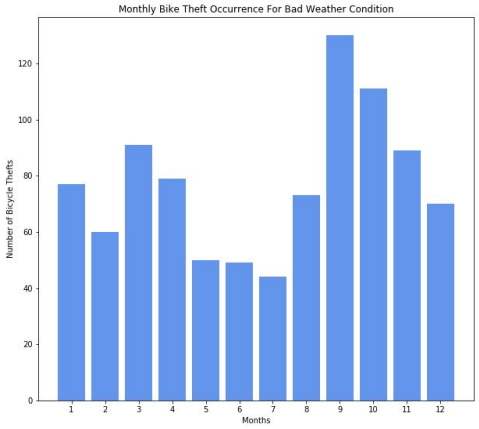
Using a scatter-plot, we can visualize the average number of bike thefts that occur every hour. There are the numbers averaged across all years of theft. The top hours where thieves strike the most are at 6:00 PM, 12:00 PM and 12:00 AM.

What type of weather condition does the theft most occurred?



One assumption that we want to explore is that the amount of bike theft occurring in good weather conditions is higher than the amount of bike theft occurring in bad weather conditions. We want to explore this as bad weather condition hinders a person's ability to ride their bike as well as people tend to use transit or drive instead. By plotting the distribution of crime for each type of weather using the bar graph, we can see that the number of thefts occurring under good weather conditions (Mainly Clear, Mostly Cloudy, Cloudy, Clear) are significantly higher than those occurring under bad weather conditions (Rain, Rain Showers, etc.)

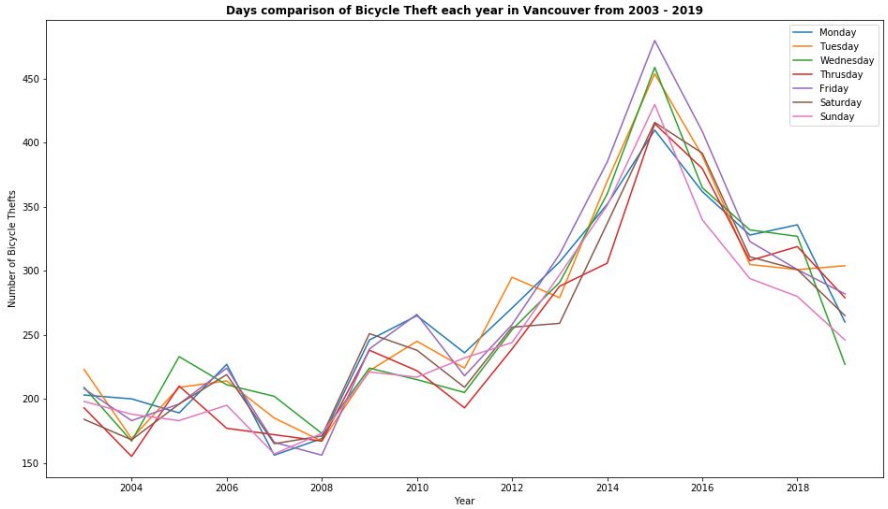




Subsequently, we started plotting the monthly distribution of the thefts for good and bad weather to determine whether the theft occurs more in good than bad weather conditions (left and right respectively). First, we need to determine if both graphs are normal to perform the test. It turns out that both of these are normal with the good and bad weather having a p-value of 0.1718 and 0.4753 which are greater than 0.05. By conducting a one-way t-test, we concluded

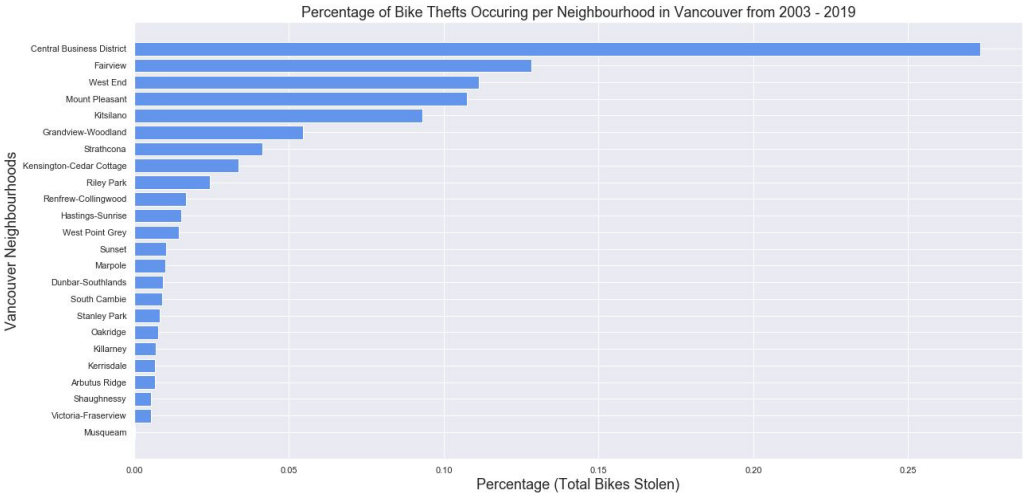
that the mean of crime for good weather is greater than the mean of crime for bad weather since the p-value < 0.1(=0.05*2), and the t-statistic > 0

Do Days Of The Week Affect Bicycle Theft?



Another assumption we have is that bike theft has no correlation with any days of the week. From the data above we can visualize that every day of the week has the same number of bikes stolen from 2003 to 2019. By grouping the data to each day of the week and use a linear plot, we can clearly see that each day does not affect when theft occurs. Using ANOVA test, we categorized the number of bike thefts each day of the week produced a p-value of 0.976. From this, we can say that the means between each day are relatively the same. In other words, the number of bike thefts for each day of the week is independent of each other.

What Neighbourhoods Are Most Vulnerable to Bike Thefts?



Vancouver has many different large neighbourhoods that differ in population. We calculated the percentage of total bikes stolen per neighbourhood. The Central Business District, also known as Downtown, has over a quarter of the total bikes stolen over the years.

2. Predictive Analysis

Can We Predict When Bicycle Theft Will Occur?

	Decision Tree	Random Forest	Naive Bayes	Gradient Boosting
Training Data	0.084	0.101	0.079	0.133
Valid Data	0.091	0.097	0.085	0.128

We want to predict the hour a bike theft will occur given the month, day, neighbourhood and hundred block. We are trying to predict a discrete target, so we will need to train a classification model. In order to use neighbourhood and hundred block as features for our model, we had to convert them into categorical numbers. Inputting text does not work as a parameter for our model, so we used label encoding for both of these features. By using four different classifiers, we found that all of the models performed relatively the same. Each model's parameters were optimized for accuracy on the test data using grid search and cross-validation.

CONCLUSION

Bike theft in Vancouver is a growing problem in Vancouver. Through our data analysis, we've found various key information to help predict bike theft. During the summer is when most thefts occur. 6:00 PM, 12:00 PM and 12:00 AM are the times when thefts occur the most. The most vulnerable neighbourhood is the Central Business District with Fairview and Westend being next. In addition, weather is also a factor for bike theft occurrence. From our analysis, we concluded that most thefts occur under good weather conditions, such as Mainly Clear, Mostly Cloudy, Cloudy, Clear, because it is the most likely time people ride their bike. If you don't want your bike to be stolen, avoid leaving your bike unattended during these areas, times and weather conditions.

Our most accurate model only predicts the correct hour when theft occurs with a 13% accuracy. One reason we think our model doesn't perform as well is that the thefts could possibly be not planned all the time. Whenever a thief sees a bike that is vulnerable, they will steal the bike. This means that the factor of vulnerability for each bike must also be included which is not possible with the current dataset. Therefore, we concluded that it is not possible to predict which hour, bike theft will occur using a machine learning model.

CHALLENGES

Since this project just focuses on specifically Vancouver, we are unable to include some of the data that we took into consideration. We considered putting in Vancouver population data. The population data did not contain the same columns in order to join the two datasets in our project. More specifically, there is no street name or hundred block name included in the population dataset, so by including the data, we could not conclude anything in particular.

While our weather data contains at least 1 weather condition information for each day, missing weather conditions for some portion of time for each day have a significant impact on our analysis since the original bike theft data contains roughly about 30,000 entries. By joining the weather and the crime data, the resulting dataset is around 3,000 entries which are 10% of the original dataset. The amount of data is enough for us to conduct the statistical analysis about whether the crime occurs more in good or bad conditions. However, it is not sufficient enough for us to attempt to create a robust prediction model as less data means less accuracy for training and validating the model.

Initially, for our predictive analysis, we were not sure if we wanted to solve a regression or classification problem. With regression, we were thinking of predicting the number of bikes stolen for a particular hundred block. If we proceeded with solving a regression problem, we were not sure how to draw the polynomial best-fit-line. We decided predicting bike theft with classification would be an easier alternative. Given the month, day, neighbourhood and hundred block, we would predict what hour the bike theft would occur. At first, we were not sure how to use classification with multiple features. In class, we only trained with one feature in each axis. Once we figured out how to train the machine learning model with multiple features, we were unsure of its accuracy. With the amount of data we had, we thought it would be higher. After thinking about what could have gone wrong, we started to think this could be correct. Bike theft could be more of an in the moment action than a planned one. In addition, there are other important factors that contributes toward the theft, such as the vulnerability factor of a bike, which we are unable to obtain.

PROJECT SUMMARY EXPERIENCE

Josh Vocal

- Trained a machine learning model to predict what hour bike theft will occur in Vancouver with a 13% accuracy using Gradient Boosted Classification.
- Visualized areas and times around Vancouver where bike theft occurs with Matplotlib and Seaborn which brought awareness to cyclists.
- Predicted yearly bike theft in Vancouver using Linear Regression which confirmed that it is a growing problem happening.

Leo Mai

- Prepared weather data by using Extract-Transform-Load methodology, Pandas and Numpy to obtain the necessary information for analysis
- Constructed visualizations by using Matplotlib to show the bike theft distribution for each weather status
- Conducted statistical analysis by using T-test to show that bike theft occur more in good weather conditions.
- Facilitated and assisted team meetings by communicating with members which ensured members of their tasks and work progress

Kevin Antonius

- Conducted statistical analysis on using T-test and ANOVA showing that Days of the week does not have any correlation with the number of bike thefts.
- Visualized the data using Matplotlib to show the bike theft groupby days which proof the uncorrelation
- Prepared weather data by using Extract-Transform-Load methodology, Pandas and Numpy to obtain the necessary information for analysis