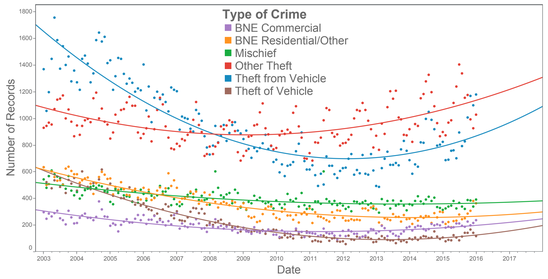
VanCrime: An analysis of Vancouver crime data

Juan Sarria, Loong Chan, Walther Maciel
Simon Fraser University, Burnaby, BC, Canada

Objectives

Crime is a concern for everyone, and since we live in Vancouver, we decided to analyze the city's crime data. In this study, we looked into finding *the safest neighbourhoods of Vancouver*, as well as *the most common offences reported in the dataset*.

With a deeper look into the data, we also wanted to find *which factors correlate the strongest with criminal activity*. And finally, we want to be able to predict *how crime rates are going to evolve in the future* within the city.



Methodology

In order to answer our questions, we relied heavily on visual analytics tools such as *Tableau* to generate insights, and on *scikit learn* + *Python* notebooks to test and generate predictive models.

Additionally, a generous amount of time was invested into data acquisition and wrangling, which was mostly done using *Pandas dataframes*.

The feature vector generated encompasses multiple location-based attributes of each one of the crime points we had. These features vary from how well illuminated the area is to the distance from the nearest transportation hubs.

In addition to location-based attributes, we also considered weather as a time-based attribute for each offence.

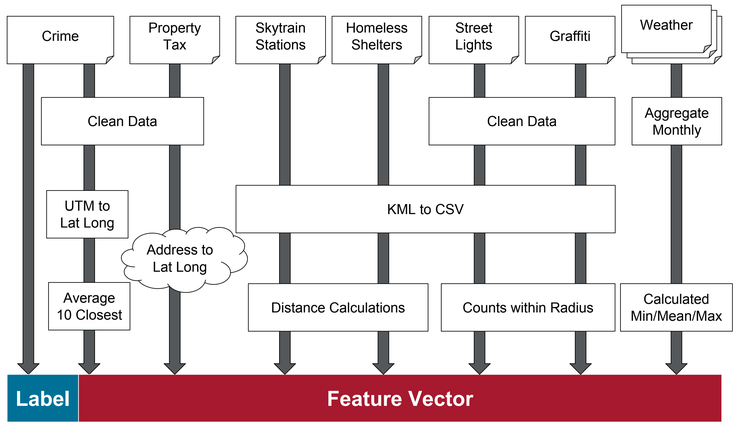
From these models, we were able to generate predictions about crimes happening in all areas of the city.

Datasets

All data was provided by the Vancouver Open Data Project and Environment Canada.

- **Crime:** Every entry represents a crime with coordinates, type, year and month.
- **Property Tax:** Provides property value, address, tax year and more.
- **SkyTrain Stations:** Provides coordinates of each SkyTrain station in the city.
- **Homeless Shelters:** Contains the coordinates of homeless shelters for men, women, youth and adults.
- **Street Lights:** Coordinates of all Vancouver light poles.
- **Graffiti:** Coordinates of graffiti identified by city staff.
- **Weather:** Daily weather reports from 3 different weather stations around Vancouver.

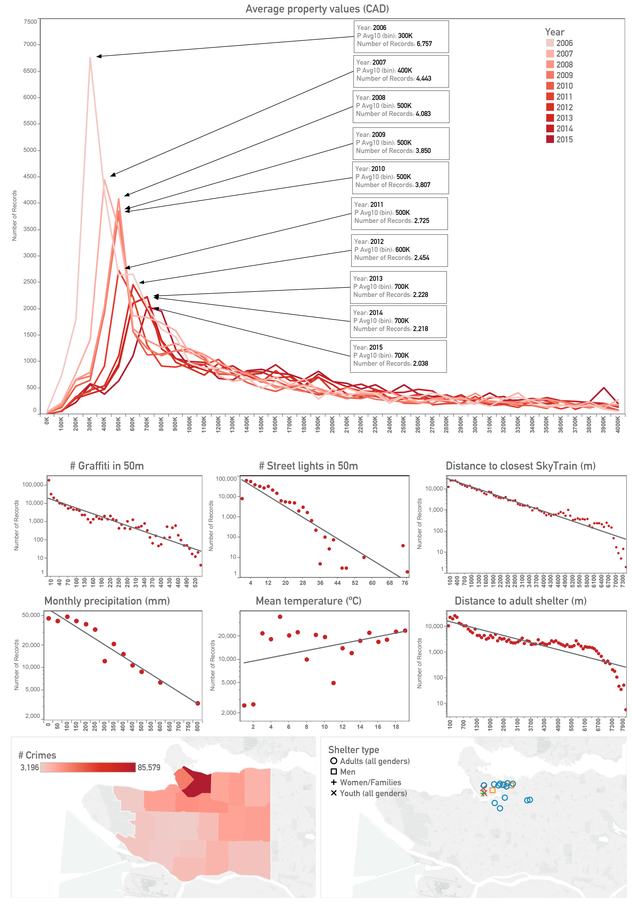
Data Pipeline



Results

To answer some of the questions in the Objectives:

- Safest neighbourhood: Shaughnessy
- Most common offence: Other Theft
- Strongest factors that correlate crime:
  - Distance from a Skytrain station
  - Monthly precipitation



With the available data, we created both a regression and classification model in order to find the expected number of crimes and the conditional probability of each type of crime respectively.
The resulting classification model has an accuracy of 49.8% and the regression model has an R^2 score of 0.435.

CMPT 733, Final Project, Professional M.Sc. in Big data, Apr/2016