Forecasting Canada Cities Crime

Student Name

Assignment #2 “Data Collection and Preparation”

INFO8136 – Descriptive Analytics

Table of Contents

[Forecasting Canada Cities Crime 1](#_Toc168433514)

[Introduction 3](#_Toc168433515)

[Metrics Parameter 3](#_Toc168433516)

[Data Source and Attributes 3](#_Toc168433517)

[Canada Ministry Dataset 3](#_Toc168433518)

[Attributes Description 3](#_Toc168433519)

[Local Boundary Data (WGS84 EPSG:4326): Open Data Portal 6](#_Toc168433520)

[Attributes Description 6](#_Toc168433521)

[Attributes Description 6](#_Toc168433522)

[References 6](#_Toc168433523)

# Introduction

The Ministry furnishes daily crime data with the aim of augmenting community awareness regarding policing activities across various crime categories. This study employs time series analysis to investigate the feasibility of accurately predicting crime occurrences. Currently, the available data for visualization via GeoDash3 is restricted to 5,000 cases and remains static. Utilizing the underlying dataset, we have delved into insights regarding crime trends, seasonality, and projected forecasts up to December 31, 2022. The dataset utilized in this report spans from January 1, 2003, to October 15, 2020, encompassing details such as the Type of Crime, Date and Time of occurrence, and location. 655,912 crimes were recorded on the VPD Geodash Open Data platform within this timeframe, containing nine features. Notably, the predominant crimes in the city, as depicted in the chart below, are Theft from Vehicle (34%), Mischief (14%), and Break and Enter Residential (11%). For neighborhood analysis, the (x, y) coordinates were converted into geometries and subsequently matched with the GeoJSON file containing neighborhood polygons sourced from the Cities Local Boundary data. This enabled accurate identification of crime locations lacking explicit neighborhood identification.

In this analysis, we will explore crime data for eight cities include Vancouver, Toronto, Ottawa, Edmonton, Ontario, Manitoba, Quebec and British Columbia, focus on various crimes. Our aim is to support the RCMP in their collaboration with cities local police. We will examine incidents such as robbery incidents, identify trends, and make predictions based on the data. This will help in proactive crime prevention.

# Metrics Parameter

# Data Source and Attributes

Canada Ministry Dataset: [www.statscan.ca](http://www.statscan.ca/), Geodash Open Data

Attributes Description:

Description: GeoDASH, short for Geographic Data Analysis and Statistics Hub, serves as a tool for mapping and analyzing crime data to keep residents informed about local crime activities. The dataset contains information about crimes that have taken place within the city, sourced from the Records Management System. It is important to note that crime classifications and file statuses are subject to change due to ongoing police investigations. The data has taken measures to safeguard the privacy of all individuals involved in reported incidents, ensuring that no personal or identifying information is disclosed in the data. For incidents involving Offences against a Person, the locations have been intentionally randomized across several blocks and offset to an intersection, with no specific time or street location provided. For property-related offences, the data offers the location to the hundred block within the general area, but all locations are offset, and users should refrain from interpreting them as related to any specific person or property.

Type:

Break and Enter Commercial: "Commercial Break and Enter" refers to the illegal act of unlawfully entering commercial premises with the intention of committing a crime.

Break and Enter Residential/Other: "Residential Burglary" involves unlawfully entering a home, apartment, garage, or similar property with the intention of committing a crime.

Vehicle Collision or Pedestrian Struck (with Fatality): Includes incidents where pedestrians or cyclists are hit and fatally injured by vehicles. It also encompasses fatal accidents involving vehicles colliding with each other, although these are less common in the overall dataset. It is important to note that there is no specific neighborhood information provided.

Vehicle Collision or Pedestrian Struck (with Injury): Includes all categories of vehicle involved accidents with injuries. This includes pedestrian and cyclist involved incidents with injuries. Note: There is no neighbourhood information

Homicide: A person, either directly or indirectly, through any method, leads to the death of another individual.

Mischief: A person engages in mischief when they intentionally cause harm, destruction, or defacement to property, which encompasses acts of public mischief directed towards others.

Offence against a Person:

An assault against an individual resulting in potential injury, involve the use of a weapon.

Other Theft: Theft involving personal belongings such as a purse, wallet, cellphone, laptop, or a bicycle, among other items.

Theft from Vehicle: Theft of property from a vehicle

Theft of Vehicle: Stealing a car, motorcycle, or any motorized vehicle

Theft of Bicycle: Stealing of a bike. Note: No specific locality details provided.

YEAR: A numerical value containing four digits denoting the year in which the documented criminal incident took place.

MONTH: A numerical entry denoting the month in which the documented criminal activity took place.

DAY: A numerical value consisting of two digits, denoting the specific day within a month when the documented criminal incident took place.

HOUR: A numeric category consisting of two digits, denoting the hour in a 24-hour format when the reported criminal incident occurred. Note: This data originates from the police inquiry. Time details will not be disclosed for crimes categorized under Offences Against a Person.

MINUTE: A numerical field comprising two digits indicating the minute of occurrence for the reported criminal activity. Note: This data is derived from police investigations. Time details will not be disclosed for crimes categorized under Offences Against a Person.

HUNDRED\_BLOCK: The generalized area where the reported criminal activity occurred.

NEIGHBOURHOOD: The information employs Statistics Canada's characterization of neighborhoods within municipalities. These neighborhoods in urban areas are delineated according to the census tract (CT) framework within a census metropolitan area (CMA).

X: Coordinate values (UTM Zone 10) indicating the location of the reported criminal activity (offset for privacy).

Coordinate values are projected in UTM Zone 10. All data offset and users should not interpret any locations as related to a specific person or specific property.

Y: Coordinate values (UTM Zone 10) indicating the location of the reported criminal activity (offset for privacy).

## Local Boundary Data (WGS84 EPSG:4326): Open Data Portal

Attributes Description:

Description: The dataset includes the delineations for the 22 local regions within the city (referred to as local planning areas). Typically, these local area borders align with street centerlines and are provided in GeoJSON format. GeoJSON is a format for exchanging geospatial data, relying on JavaScript Object Notation (JSON). It utilizes the World Geodetic System 1984 as its geographic coordinate reference system and measures distances in decimal degrees (specified by the IETF).

Attributes Description:

Type:

MAPID: A label indicating the locality (e.g., AR for Arbutus-Ridge).

Name: Neighborhood full name.

Geom: Local Area spatial representation involves depicting points, curves, and surfaces within a coordinate framework.

geo\_point\_2d: For the "Point" type, the "coordinates" attribute represents a solitary location.

# ETL Process

The data analysis process divided into two major stages:

1. Data Pre-processing

2. Data Visualization

Decisions taken at both these stages takes the analysis deliverables.

## Stage 1: Data Pre-processing

During this stage, we will load and transform the data using PowerQuery. The objective is to prepare the dataset for loading into PowerBI to generate visualizations. Given the large size of the dataset, we will filter it to retain only the variables relevant to the visualizations. The filteringbased on the three dimensions of the dataset: geography, violation, and statistics.

All selected variables are based on the expected deliverables. Ideally, a local authority would like to reduce violent crimes in a vicinity and then follow up to reduce non-violent crimes. Hence, this analysis takes all violent criminal code violations for the analysis.

Finally, to determine the top eight cities with the highest crime in each cities, we use python and MYSql code. A python function carry out data manipulation to get the top N cities for the defined geography, violation and statistic.

## Stage 2: Data Visualization

This stage create different visualizations to communicate the objectives defined by the team.

Key Findings