Unveiling Toronto's Major Crime Indicators:

A Comprehensive Analysis across Temporal, Geographic, and Pattern Dimensions

Amos Syh Ern Chew: n01533575
Ricardo Joaquin Hornedo Aldeco: n01538048
Wenhao Fang: n01555914
Data: 2/12/2023

1. Abstract

This project employs big data technologies, including Hadoop, Hive, Spark, and Zeppelin, to analyze major crimes in Toronto from 2014 onward. Utilizing the authoritative Major Crime Indicator dataset, we implemented a logical model, crime_view, for efficient analysis. Temporal insights highlight dynamic crime trends, while geographic analysis identifies the top 5 regions with both high and low crime numbers. Pattern analysis reveals assault as the predominant crime category and provides insights into crime distribution across different premises. This study introduces practical business applications regarding urban safety in Toronto.

2. Data Source and Logical Model Design

2.1. Data Source

Major Crime Indicator (https://open.toronto.ca/dataset/major-crime-indicators/) is the dataset about crime occurrences provided by the city of Toronto's open data. It has the following features:

- It is an authoritative data source Published by Toronto Police Services.
- It includes categories of major crimes in Toronto.
- It contains historical data from 2014.

Precisely, our dataset is a CSV file that has 323296 rows of major crime data across 158 City of Toronto neighborhoods.

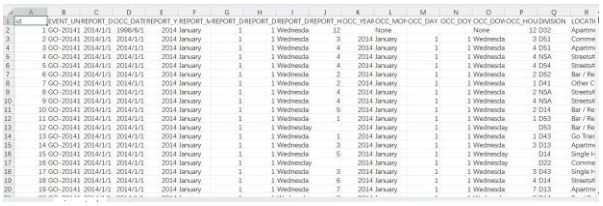

***Figure 1**. CSV file*

2.2. Logical Model Design

For ease of analysis, we selected 7 out of 27 columns to build our logical model and designed our **crime_view** model as follows.

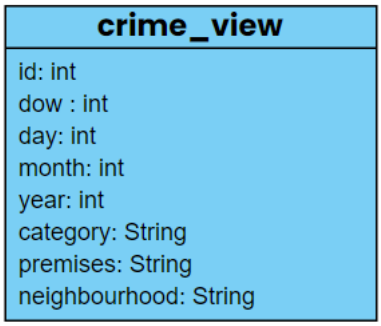

***Figure 2**. Logical Model*

3. The Implementation of the Logical Model

To implement our warehouse model, we leverage ***Hadoop***, a powerful big data technology, to harness the benefits of distributed storage and processing. Placing our CSV file into Hadoop ensures scalability, fault tolerance, and efficient data handling.

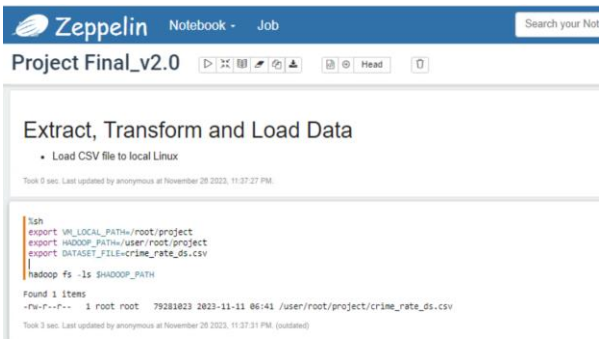

***Figure 3**. Load CSV file (Hadoop)*

Subsequently, we employed ***Hive***, another big data technology, to establish a robust database infrastructure, including the creation of the **crime_rate_ext** external table, serving as a direct reference to the original CSV file, and the **crime_rate_orc** internal table, facilitating data importation, transformation, and staging within Hive. The pivotal **crime_view** serves as a central viewpoint for in-depth analysis.

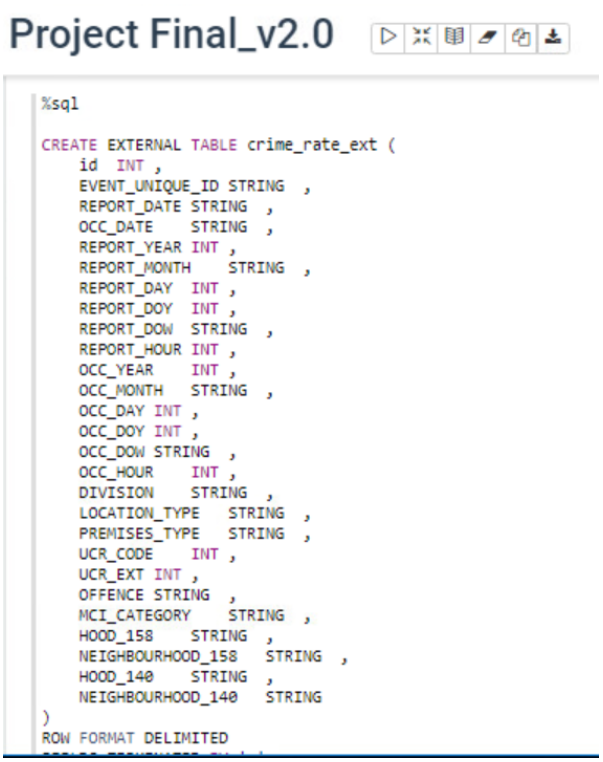

Figure 4. *Creation of crime_rate_ext (Hive)*

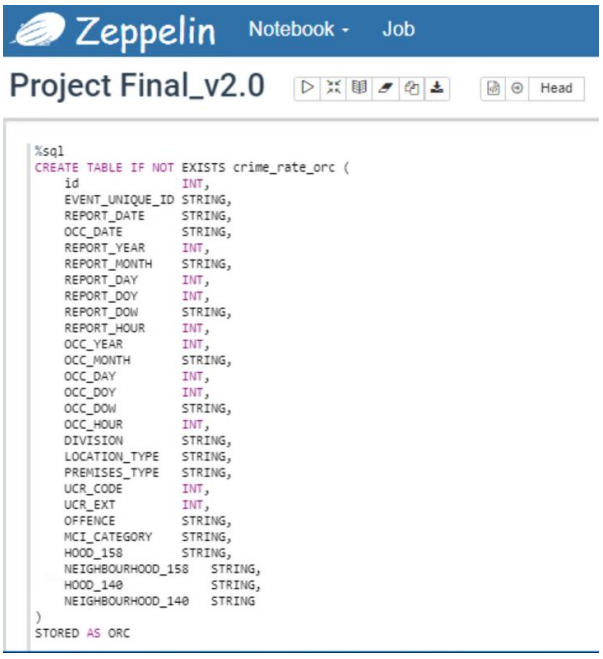

Figure 5. *Creation of crime_rate_orc (Hive)*

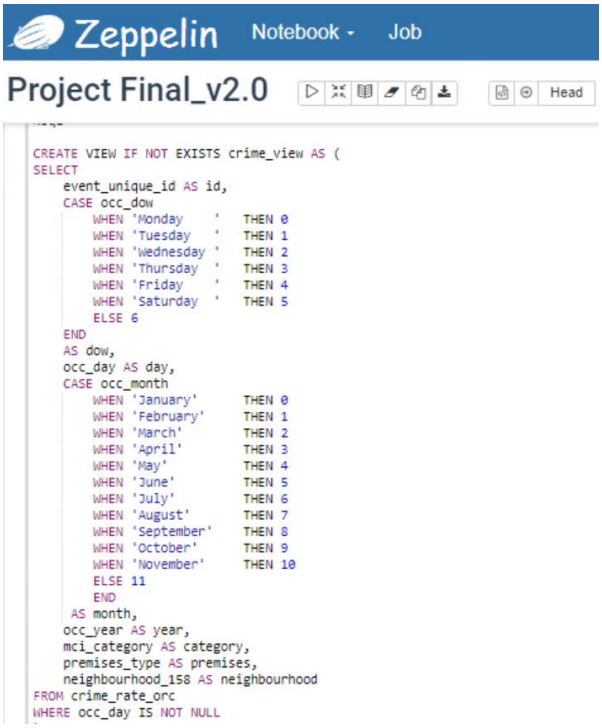

Figure 6. *Creation of crime_view (Hive)*

Zeppelin Notebook acts as the intuitive interface, providing a seamless platform for command input and execution, thereby streamlining the entire data processing and analytical workflow. This cohesive integration of ***Hadoop***, ***Hive***, and ***Zeppelin*** optimally positions our warehouse model for effective and scalable crime data analysis.

4. Temporal Analysis

4.1. General Trend

To grasp an overarching view of crime rates, we initiated a temporal analysis by calculating

the annual count of criminal incidents in Toronto from 2014 onwards. Utilizing Zeppelin, we performed SQL queries and generated trend charts to visualize the annual crime patterns.

```
%sql
SELECT
    year AS Year,
    COUNT(*) AS Crime_count
FROM crime_view
WHERE year >= 2014
GROUP BY year
ORDER BY year ASC
```

***Code 1**. Crime Count Yearly*

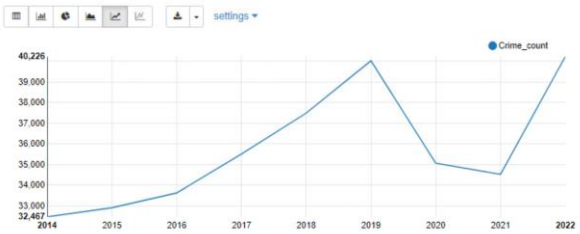

***Figure 7.** Crime Count Yearly*

The line chart depicting Toronto's annual crime trends from 2014 to 2022 reveals dynamic fluctuations. Starting at 32,467 incidents in 2014, the crime rate steadily increased to a peak of 40,026 in 2019. Subsequent years witnessed a decline in 2020 and 2021, followed by a sudden upturn to 40,225 incidents in 2022.

4.2. Day of Week

Analyzing the crime patterns by day of the week, we executed a query to compute the total number of crimes for each day and visualized the findings in a line chart.

```
%sql
SELECT
    dow AS Day_of_Week,
    CASE dow
        WHEN 0 THEN 'Mon'
        WHEN 1 THEN 'Tue'
        WHEN 2 THEN 'Wed'
        WHEN 3 THEN 'Thur'
        WHEN 4 THEN 'Fri'
        WHEN 5 THEN 'Sat'
        ELSE 'Sun'
    END
    AS d_o_w,
    COUNT(*) AS Crime_count
FROM crime_view
GROUP BY 1
ORDER BY 1 ASC
```

***Code 2**. Day of week*

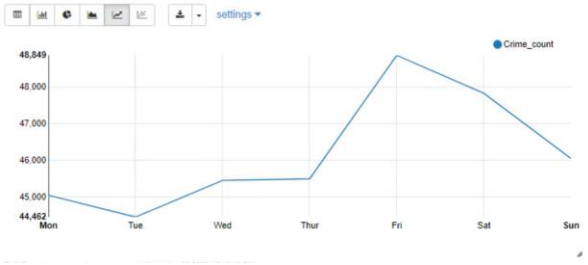

***Figure 8**. Day of week*

As depicted in the line chart, crime counts on Friday, Saturday, and Sunday exhibit a slight elevation compared to other days of the week. Specifically, Fridays consistently have the highest numbers, while Tuesdays consistently have the lowest figures.

5. Geographic Analysis

5.1. Top 5 Region with Lowest Crime

'Location, location, location.' - A mantra in the real estate industry strengthens the impact of the geographical factor. In this section, we analyze the pivotal role of location on the regional distribution of crime in Toronto City. Employing ***Spark*** in ***Zeppelin***, we queried the top 5 regions with the lowest crime numbers, providing valuable perspectives on safer areas.

```
%spark2

val df_low_region = spark.sql("SELECT
COUNT(*) AS crime_count, neighbourhood
FROM crime_view GROUP BY neighbourhood
ORDER BY crime_count ASC LIMIT 5")

df_low_region.show()
```

Code 3. *Regions with low crime*

```
+-----------+------------------+
|crime_count|     neighbourhood|
+-----------+------------------+
|        519|Lambton Baby Point|
|        591|  Woodbine-Lumsden|
|        595|         Guildwood|
|        617|        Maple Leaf|
|        642|    Yonge-St.Clair|
+-----------+------------------+
```

Table 1. *Regions with low crime*

As the output is shown below, the neighborhood Lambton Baby Point has the lowest crime number, 519, followed by Woodbine-Lumsden, Guildwood, Maple Leaf, and Yonge-St.Clair.

5.2. Top 5 Region with Highest Crime

Leveraging ***Spark***, we conducted the query for the top 5 regions with the highest crime numbers, uncovering the dangerous neighborhoods in Toronto City.

```
%spark2

val df_highest_region =
spark.sql("SELECT COUNT(*) AS
crime_count, neighbourhood FROM
crime_view GROUP BY neighbourhood
ORDER BY crime_count DESC LIMIT 5")

df_highest_region.show()
```

Code 4. *Regions with high crime*

```
+-----------+--------------------+
|crime_count|       neighbourhood|
+-----------+--------------------+
|       8803|West Humber-Clair...|
|       7746|           Moss Park|
|       6840| Downtown Yonge East|
|       6495|  Yonge-Bay Corridor|
|       6236|    Wellington Place|
+-----------+--------------------+
```

Table 2. *Regions with high crime*

As the output is shown below, the neighborhood West Humber-Clair has the highest crime number, 8803, followed by Moss Park, Downtown Yonge East, Yonge-Bay Corridor, and Wellington Place. We notice that

Humber College (North Campus) is located within the West Humber-Clair neighborhood that has the highest crime number.

***Figure 9**. Categories*

6. Pattern Analysis

After conducting a comprehensive analysis of the data in both temporal and spatial dimensions, the next step involves delving into pattern analysis, encompassing an examination of crime categories and locations.

6.1. Crime Category Analysis

Leveraging ***Hive*** technology, we queried crime numbers based on different categories, revealing a distinctive pattern in Toronto. Assault emerges as the primary crime, accounting for 54% of incidents, trailed by Break and Enter at 19%, and Auto Theft at 14%.

```
%sql
SELECT
    category,
    COUNT(*) AS crime_count
FROM crime_view
GROUP BY
    category
ORDER BY
    crime_count
```

***Code 5**. Categories*

Furthermore, we queried to calculate a year-to-year crime number based on different categories and employed a line chart for visualization. As shown below, all types of crime numbers rise in 2022, and auto theft has increased significantly in recent years, exceeding the Break and Entry in 2021 and becoming the second major crime.

```
%sql
SELECT
    COUNT(*) AS crime_count,
    category,
    year
FROM crime_view
WHERE year >= 2014
GROUP BY
    year, category
ORDER BY
    crime_count, category DESC
```

***Code 6**. Categories year-to-year*

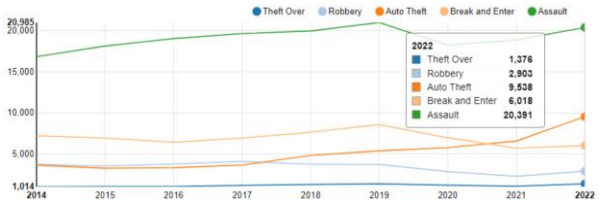

***Figure 10**. Categories year-to-year*

6.2. Premises Categories

Finally, we performed a Hive query against crime numbers based on different premises.

```
%sql

SELECT
    premises,
    COUNT(*) AS crime_count
FROM crime_view
GROUP BY
    premises
ORDER BY
    crime_count
```

***Code 7**. Premises*

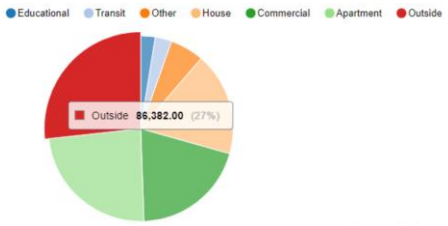

***Figure 11**. Premises*

The analysis result indicates a distribution in crime locations, with incidents occurring outside comprising 27%, closely followed by apartments at 24%, commercial areas at 20%, and houses at 18%.

In addition, a query on year-to-year premises unveiled a significant surge in incidents for outside locations, commercial areas, and houses, particularly in 2022.

```
%sql
SELECT
    COUNT(*) AS crime_count,
    premises,
    year
FROM crime_view
WHERE year >= 2014
GROUP BY
    year, premises
ORDER BY
    crime_count, premises DESC
```

***Code 8**. Premises*

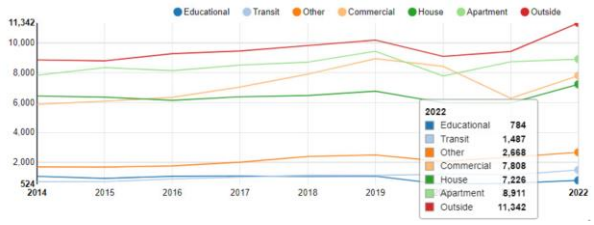

***Figure 12**. Premises year-to-year*

7. Practical utilization

Our analysis of major crime in Toronto City can be utilized in:

- **Law Enforcement Resource Optimization**: Direct law enforcement resources to regions identified with high crime numbers, ensuring a targeted and effective approach to crime reduction.

- **Housing Market Considerations:** Leverage crime analysis insights as a key determinant in housing market evaluations, emphasizing the safety of neighborhoods as a critical factor for potential homebuyers.
- **Tailored Premises Security Strategies:**

  Develop premises-specific security strategies based on the analysis, acknowledging the varying crime numbers associated with different premises types for a more nuanced and effective security approach.

8. Conclusion

In this project, we utilized big data technologies, including ***Hadoop***, ***Hive***, ***Spark***, and ***Zeppelin*** to analyze the major crimes in Toronto City.

Within the temporal analysis, we found that Annual crime trends showcased dynamic fluctuations from 2014 to 2022. The day-of-week analysis revealed distinct weekly patterns, with Fridays consistently having the highest crime counts.

Regarding geographic analysis, we identified both the top 5 regions with the highest and lowest crime numbers in Toronto. The Humber College (north campus) is in the highest crime region.

As to pattern analysis, crime category analysis highlighted assault as the major crime, comprising 54% of crime. Premises analysis showcased the distribution of crime locations, with incidents occurring outside, in apartments, commercial areas, and houses.

Our analysis can serve as a strategic guide, empowering law enforcement with targeted resource allocation, informing housing market decisions through safety considerations, and facilitating the development of nuanced security strategies tailored to specific premises types.