Vancouver Crime Data Warehousing and Geo-Visualization

CMPT740
Database Systems
Course Project

Syed Tanveer Jishan
301279717

Yiji Wang
301286922

Supervised by
Prof. Wo-shun Luk

April 20, 2016

(This project has equal contribution from both the authors)

Abstract

Crime data analysis is fundamental to understanding crime patterns which will aid in preventing future crimes to happen. In this project we have build a data warehouse of Vancouver crime data which will allow users to make precise queries. Furthermore, we have provided an interactive geo-visualization system for better understanding of crime situation. We have also commenced analysis on this data to answer some of the important questions related to crime in Vancouver.

Contents

1 Introduction

Recently many governments have undertaken the open data initiative, releasing datasets to the public. People now can have a better community awareness of policing activity in Vancouver by analyzing the data for different aspects. In this project, we did analysis on the crime data provided by the Vancouver government along with various related datasets such as the census data. Aiming to find the crime situation for Vancouver neighborhoods for the past 12 years, we applied data cleaning using Extract-Transform-Load(ETL) on the data collected, built multidimensional data cube with SQL Server Analytic Services and did visualization with Tableau, a popular visualization software that provides the nice interfaces for viewing the data. Any non-expert user can easily query the cube and generate fancy graphs by doing pulling and dragging using Tableau. Apart from the visualization, we have answered some of the interesting questions related to crime in Vancouver, such as, is crime increasing in Vancouver? If so, what are the factors affecting that. Are rich neighborhoods safer? Does change in seasons has any correlation with crimes happening?

The rest of the report is divided as follows, in the Section 2 we will briefly discuss about the works that provided us with the intuition to develop our own system. In Section 3, at first we will discuss about the system pipeline and then we will described the data collection,cleaning,warehousing and visualization building process. In the Section 3 we will provide detailed answers to some of the questions mentioned in this section. Finally, we will draw conclusion and provide insights of the future work.

The rest of the report is divided as follows, in the Section 2 at first we will discuss about the system pipeline and then we will described the data collection,cleaning,warehousing and visualization building process. In the Section 3 we will provide detailed answers to some of the questions mentioned in this section. Finally, we will draw conclusion and provide insights of the future work.

2 Related Works

Vancouver Police Department[1] provides a map to visualization the crime happening around Vancouver. The map provides the almost precise geolocation of the crime event. Unfortunately the map is not interactive and do not provide scope to analyze crime data. In our system we are providing the users with the scope to slice, dice and analyze the Vancouver crime situation

Clancey[2] discusses about a system which provides crime data analysis for the New South Wales, Australia. The system gives geo-visualization accompanying with temporal trends to better understand the crime events.

Zhang[3] provided several perspective on how a crime data visualization user interface may look like. Our system follows a similar approach for the user interface provided but with more focus on geo-visualization.

3 System Implementation

3.1 System Pipeline

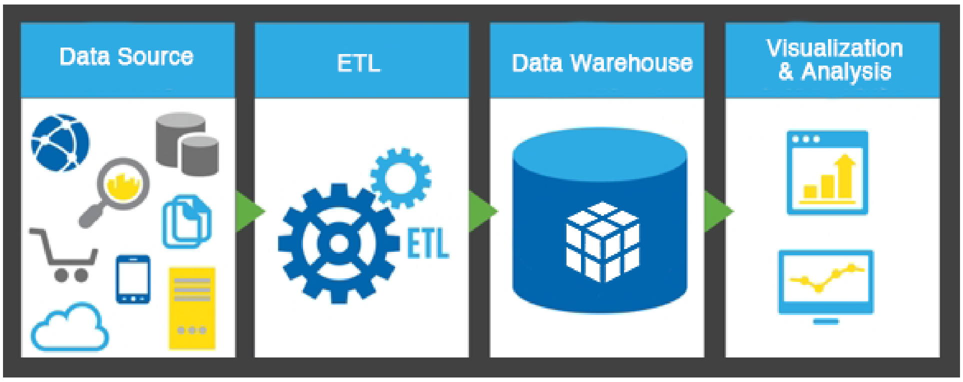


Figure 1: *System pipeline of this project.*

Above is the system pipeline of this project. First We collected the data from different official sources. Then we doing data cleaning using ETL provided by SSAS and built multidimensional cube with hierarchical structure. Finally we used commercial software Tableau for querying the cube and offer nice visualization. Based on the different charts generated from the cube, we can find some interesting and meaningful information about the crime situation in vancouver neighborhoods.

3.2 Data Collection and Data cleaning

3.2.1 Data Collection

There are 3 main types of data in excel file we collected. (1) Crime Data (2) Census Data and (3) Neighborhood polygon map data
Main Data Source: We get all data from the Vancouver official website: data.vancouver.ca
Data Description:

- Crime Data
  The crime data is an excel file with over 500K records containing the type, year and the location of the crimes from 2003 to 2015. The crimes are catalogued into 8 types: BNE(Break and Enter) Commercial, BNE Residential/Other, Homicide, Mischief, Offence Against a Person, Other Theft, Theft from Vehicle, Theft of Vehicle.

- Census Data
  The census is Canadas largest and most comprehensive data source

conducted by Statistics Canada every five years. The Census of Population collects demographics and linguistic information on every man, woman and child living in Canada with a multidimensional table for every neighborhood.

- Neighborhood polygon map data
  The neighborhood polygon map data is a data set in SHP format that contains the boundaries for the Citys 22 local areas.

3.2.2 Data Cleaning

To merge the data from different sources and filter out the information we need, data cleaning is crucial. In this project, we adopted ETL technique provided by SSAS to do the data cleaning.
Extract-Transform-Load(ETL) Extract, Transform and Load (ETL) refers to a process in database usage and especially in data warehousing that:

- Extracts data from homogeneous or heterogeneous data sources
- Transforms the data for storing it in the proper format or structure for the purposes of querying and analysis
- Loads it into the final target (data warehouse)

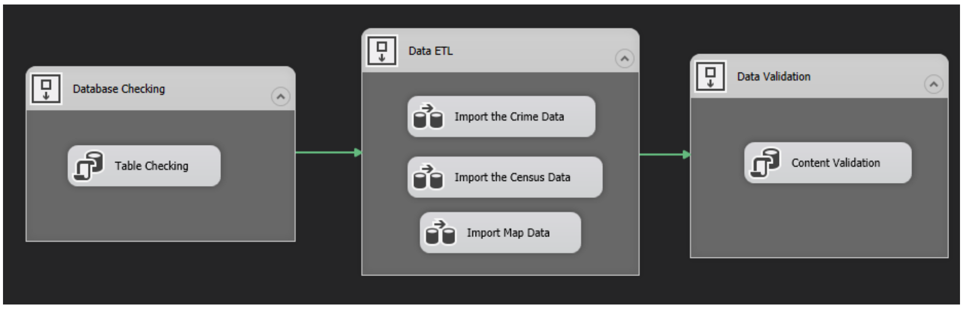


Figure 2: *This figure shows how ETL performs data cleaning. First the tables in the database will be checked to prevent from duplication import and null value. Then data will be extracted, transformed and loaded to the database. Finally the imported data will be validated.*

i Extract
In the extract step, we extracted the data from different sources into the database. This process is simple and fast since all the data are local data in excel format.

ii Transform
This step occupied 2/3 of time for the whole ETL process. There are 3 main data issue involved in this part. They are: data incompleteness, data inconsistency and data duplication.

- Data Incompleteness
  We found there are some null values in the data and also for a certain type of the crime such as homicide, the location is marked as privacy. We filtered those data because they basically show nothing for the crime in the Vancouver neighborhoods.
- Data Inconsistency
  Due to the fact that the data is collected from multiple sources, it is crucial to ensure data consistency for building the cube. We applied data matching and validation on the data to make sure the attribute names are matched correctly and the data types are the same.
- Data Duplication
  The source data have no duplication since it's officially provided. However, to retrieve data from some certain attributes such as the crime types, we need to handle the data duplication because there are over 90K rows for each type of the crimes. Here in ETL, we applied the sort function which also provided the duplication removal.

iii Load
After the previous steps, we loaded the cleaned data to the database, creating the tables required for the cube.

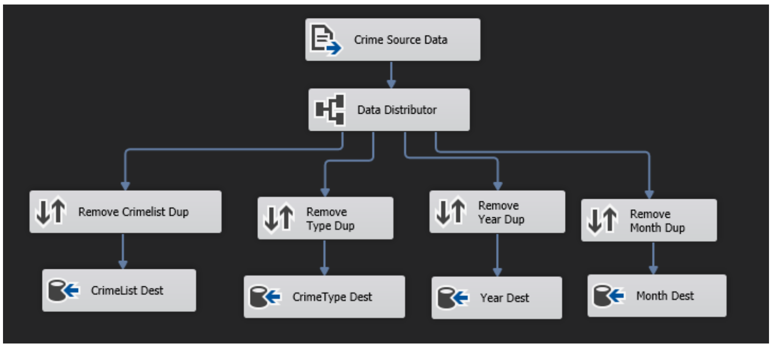


Figure 3: *This figure shows sample ETL for crime data.*

3.3 Data Warehousing

3.3.1 Database Schema

After cleaning and preprocessing the data, tables were generated using the ETL tool The design of this database is based on star schema.

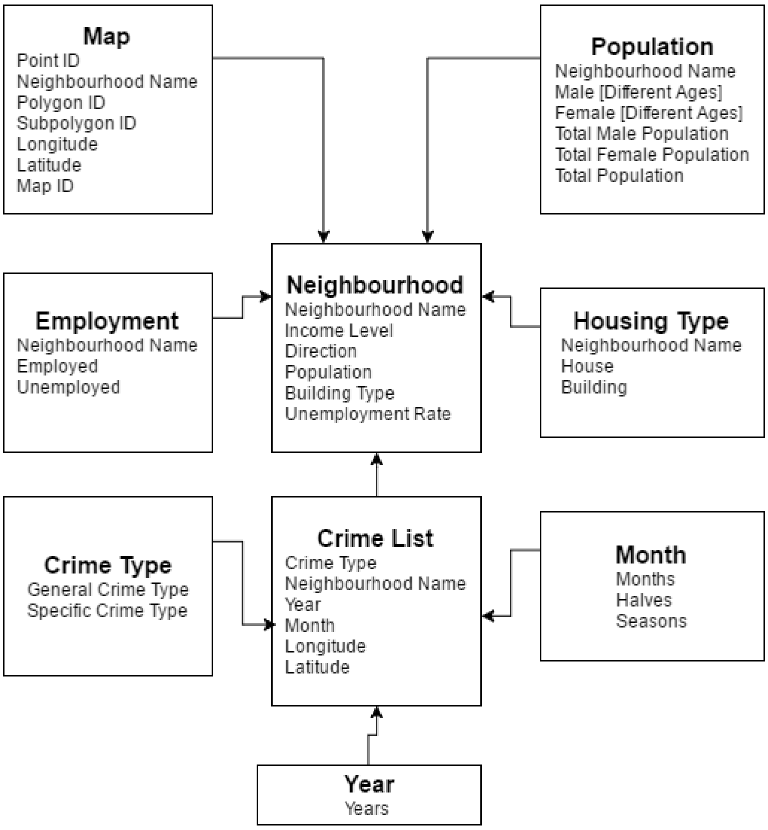


Figure 4: *This figure shows the database schema.*

In total nine tables were generated and each of them were linked using foreign keys. Neighborhood table contains the neighborhood names and every other attributes defines various aspect of the neighborhoods. For instance, the attribute Income Level indicates the median income level of people in each and every neighborhoods in Vancouver. Likewise, Direction represents in which side of the particular neighborhood belongs to. As mentioned in the previous section, population of the city is binned into three groups, high, medium and low. We also considered the building types in each of the neighborhoods. For example, some neighborhoods most contains houses while others are full of buildings. We took this factor into consideration since types of buildings in each neighborhood might have correlation with breaking and entering of residence and commercial properties. Unemployment rate is also one of the factor taken into consideration. Discretization of these columns in the Neighborhood table is aided by the Population, Employment and Housing Type tables. Furthermore, we have used a Map table which contains map designing information of the neighborhoods of Vancouver.

3.3.2 Multidimensional Hierarchical Cube

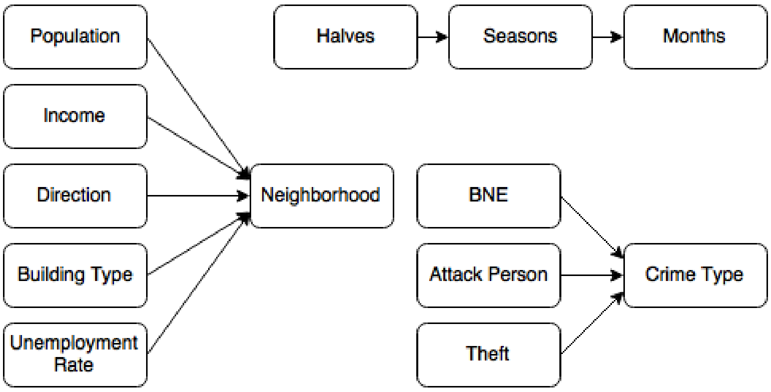


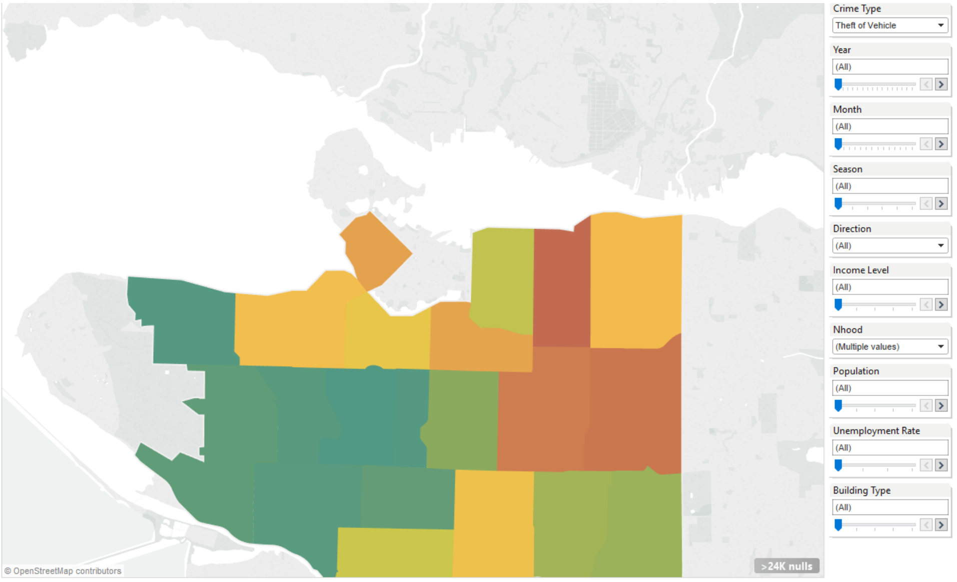
Figure 5: *This figure shows 3 types of hierarchical dimensions.*

As previously mentioned one of the goal of the project is to design a system which can answer different queries based on crimes in Vancouver neighborhoods. In order to facilitate that it is important to provide drilling down, slicing and dicing features. Therefore, a multidimensional hyper-cube was built which can provide crime count based on the following axes, neighborhood, crime type, year and month. Each dimensions except the year contains multiple hierarchies, like for example, the dimension month contains three-level hierarchies as follows, halves-¿seasons-¿months. The fact table of the cube was built using the Crime List table and different dimensions are provided by the Crime Type, Month, Year and Neighborhood table as shown in the table. All the attributes of these dimension tables contributes to the hierarchy levels of the dimensions in the cube. These multi-level hierarchies on different dimensions allow the system to answer the users question. For example, it is possible to answer which neighborhood from the west side of Vancouver where unemployment rate is low has higher number of commercial and residential property crime happening during the summers of late 2000s.

3.4 Visualization Implementation

3.4.1 Polygon Map Construction

By default Tableau do not support neighborhood level visualization. In order to provide map for neighborhoods in Vancouver we had to take into consideration the GPS coordinates defining the boundary of each neighborhood in order to create a polygon for each neighborhood. The GPS coordinates for each neighborhood was provided by the City of Vancouver in terms of shape file. However Tableau do not support shape file , therefore we had to use a utility program to convert the shape file into csv which can be easily added to the relational database. The Map table in the database is used to create geo-visualization in Tableau. In the table, Point ID is the primary key for each coordinates consisting of Longitude and Latitude. Polygon ID and Subpolygon ID represents the neighborhood polygon shapes.



Crime Type: Theft of Vehicle
Year: (All)
Month: (All)
Season: (All)
Direction: (All)
Income Level: (All)
Nhood: (Multiple values)
Population: (All)
Unemployment Rate: (All)
Building Type: (All)

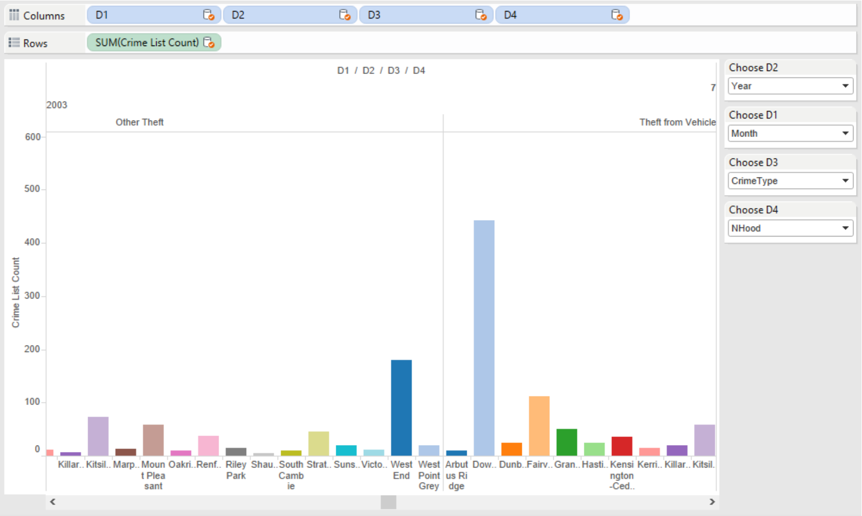
Figure 6: *This figure shows main map user interface. User can easily query the cube by choosing different hierarchical dimensions.*

3.4.2 Systems Using Calculated Field

If the data doesn't include all of the fields required to answer some questions, it is possible to create new fields in Tableau and then save them as part of the data source. For instance, in our data we have the population information for each neighbourhoods and we can calculate the crime frequency of those neighbourhoods as well. However, calculated field will helps us to solve this problem by writing the formula to take the ratio. We can create calculated fields in Tableau by defining a formula that is based on existing fields and other calculated fields, using standard functions and operators.

3.4.2.1 Multidimensional User Interface

Tableau provides the user with Drag and Drop functions to produce nice visualization. However, for the users without much experience in Database System, all they want is a menu where they can query the data by simple clicks. Therefore, we provided a multidimensional user interface using the calculated field function integrated in Tableau. User can choose the dimension they want easy by selecting in the dropdown list.



Columns: D1, D2, D3, D4

Rows: SUM(Crime List Count)

Choose D2: Year

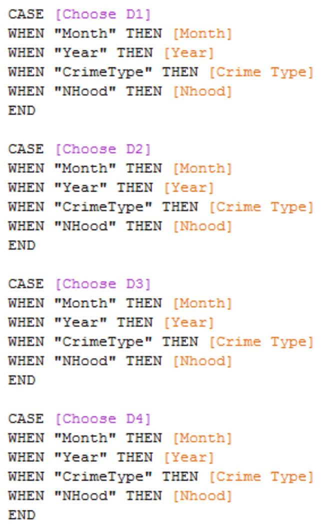
Choose D1: Month

Choose D3: CrimeType

Choose D4: NHood

Figure 7: *This figure shows the multidimensional user interface.*

With the help of calculated field funciton, user can change the dimensions dynamically to query different level of hierarchical data.



```
CASE [Choose D1]
WHEN "Month" THEN [Month]
WHEN "Year" THEN [Year]
WHEN "CrimeType" THEN [Crime Type]
WHEN "NHood" THEN [Nhood]
END

CASE [Choose D2]
WHEN "Month" THEN [Month]
WHEN "Year" THEN [Year]
WHEN "CrimeType" THEN [Crime Type]
WHEN "NHood" THEN [Nhood]
END

CASE [Choose D3]
WHEN "Month" THEN [Month]
WHEN "Year" THEN [Year]
WHEN "CrimeType" THEN [Crime Type]
WHEN "NHood" THEN [Nhood]
END

CASE [Choose D4]
WHEN "Month" THEN [Month]
WHEN "Year" THEN [Year]
WHEN "CrimeType" THEN [Crime Type]
WHEN "NHood" THEN [Nhood]
END
```

Figure 8: *This figure shows the Calculated Filed function used for the multidimensional user interface.*

3.4.2.2 Ranking System

People like to see ranking a lot. In the user interface, we provide the ranking system to show the ranks of the result for different queries. For example, The user can see the topN neighbourhood among all the neighbourhoods for all types of data for all 12 years. If more filters applied, such as the time and crime type, the ranking will change automatically. Beside finding the top N results, the user can have the full information for the ranking, i.e. see the intermediate ranknig for the neighbourhoods.

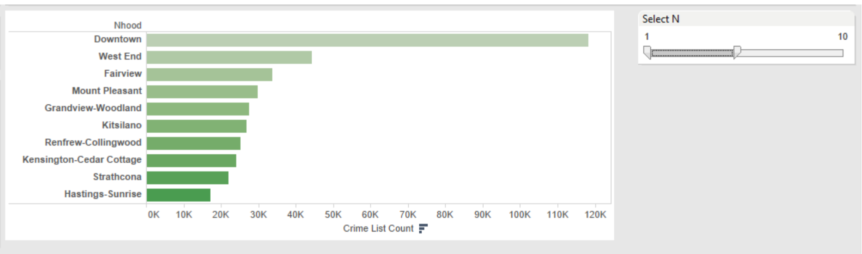


Figure 9: *This figure shows the Ranking System provided.*

3.4.2.3 Ratio on Hierarchies

It is possible to make comparative measurements among different hierarchies. For instance, in this figure above we have calculated the ratio of crime in places where population is high, medium and low. This will aid us in determining how many times more crime happens in highly populated areas compared to lowly populated areas, for this case it is almost 5 times more; similarly other values can also be taken into consideration. In the figure below, the calculation members to create the ratio on hierarchies is given below.

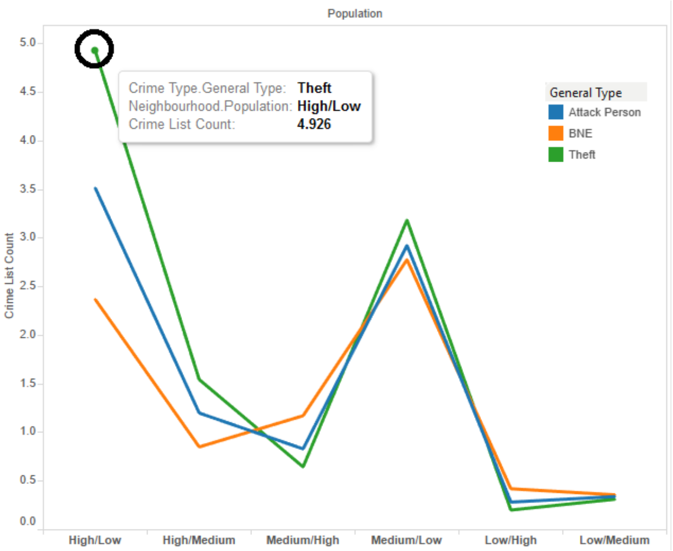
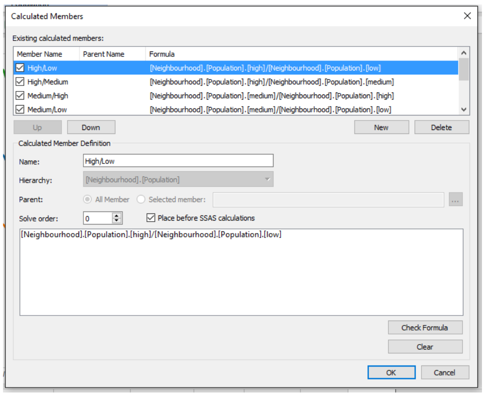


Figure 10: *This figure shows the ratio line chart for comparing neighborhoods with different population.*



Existing calculated members:

Member Name	Parent Name	Formula
☑ High/Low		[Neighbourhood].[Population].[high]/[Neighbourhood].[Population].[low]
☑ High/Medium		[Neighbourhood].[Population].[high]/[Neighbourhood].[Population].[medium]
☑ Medium/High		[Neighbourhood].[Population].[medium]/[Neighbourhood].[Population].[high]
☑ Medium/Low		[Neighbourhood].[Population].[medium]/[Neighbourhood].[Population].[low]

Calculated Member Definition

Name: High/Low

Hierarchy: [Neighbourhood].[Population]

Parent: ● All Member ○ Selected member:

Solve order: 0 ☑ Place before SSAS calculations

```
[Neighbourhood].[Population].[high]/[Neighbourhood].[Population].[low]
```

Figure 11: *This figure shows the calculated filed used to produce the ratio chart.*

4 Data Analysis

In this section, we are going to discuss about the significant insights gathered through the crime data analysis of the Vancouver neighborhoods.

4.1 Is the overall crime increasing in Vancouver?

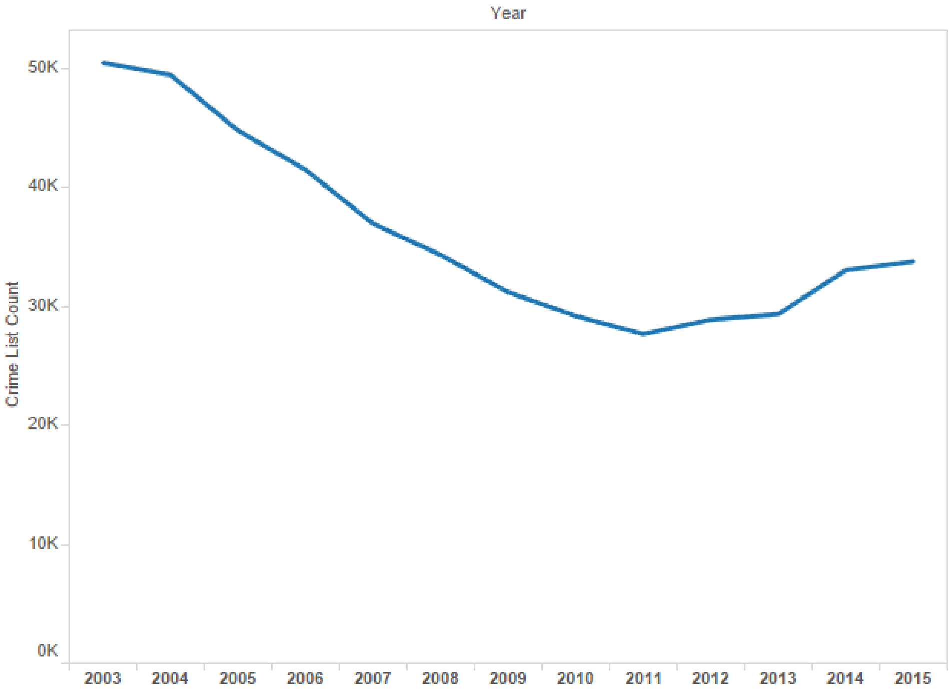


Figure 12: *This figure shows the overall crime rate from 2003 to 2015 in Vancouver.*

The figure above represents the count of all types of crimes happening in Vancouver from 2003 to 2015. From the figure, it is quite clear that the rate of crime was decreasing fairly at a steady rate until 2010. Unfortunately, since 2011 the crime is on the rise again. As a result the crime rate in 2015 is almost as high as it was 10 years back. However, the crime rate in 2015 did not increase as significantly as it did in other years since 2011.

4.2 Is it increasing with the same proportion in every neighborhoods?

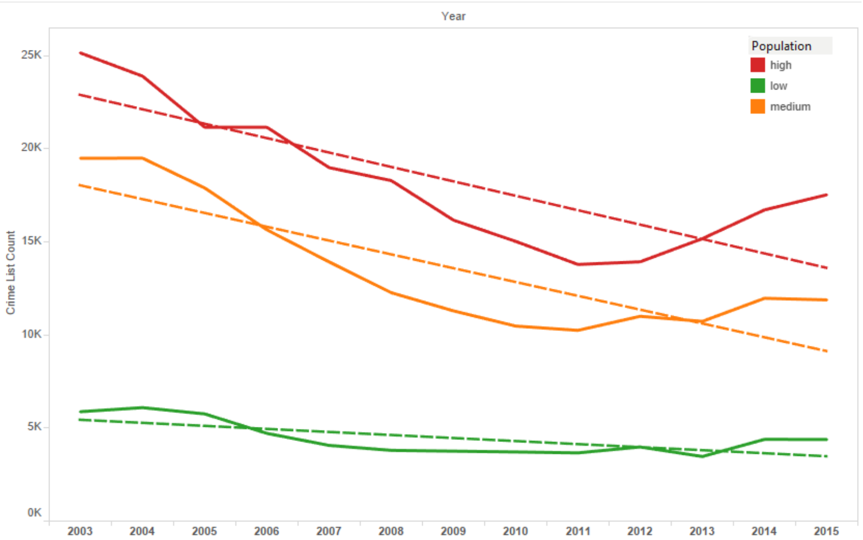


Figure 13: *This figure shows the crime rate in neighborhoods with differnet population from 2003 to 2015 in Vancouver.*

Although the crime is increasing in Vancouver since 2011, however it is not increasing at the same rate everywhere in Vancouver. In the figure above which shows the crime rate per year in Vancouver each of the lines represents the neighborhoods with different population brackets. The red line represents the neighborhood where population is high, orange represents medium and green represents low respectively. In this figure we can see that the crime rate fluctuated in the highly populated area compared to lowly populated area where it is almost steady over the years. This is quite evident that the crime is strongly correlated with the population of the neighborhoods in Vancouver. Furthermore, this gives us an intuition that it is the highly populated area which contributes to the overall increase in crime in Vancouver. It is visible in the graph that until 2010 both the highly populated and medium populated neighborhoods altogether saw sharp decrease in crime rate and a decent increase in crime rate since 2011.

4.3 What kind of crime is actually increasing?

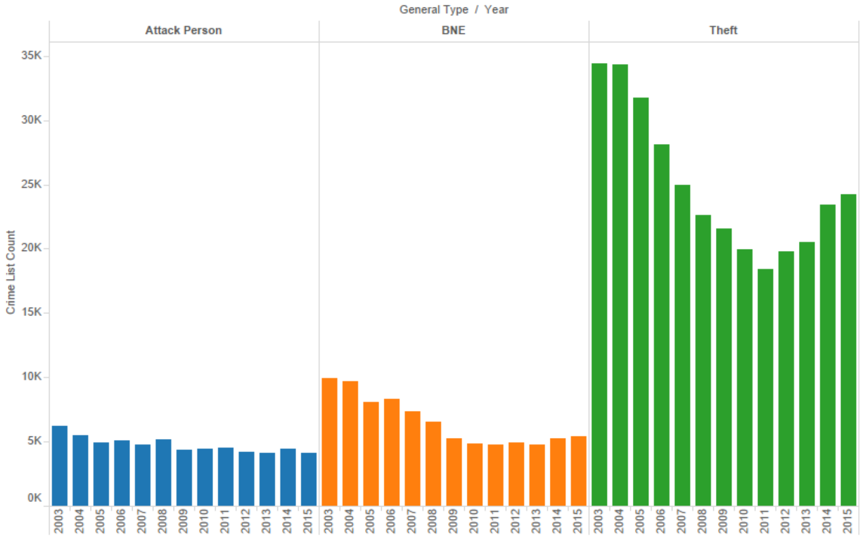


Figure 14: *This figure shows the crime rate for a certain type of crime from 2003 to 2015 in Vancouver.*

Another factor that contributes to the increase in crime among Vancouver neighborhoods is the type of the crime. From the figure above we can see that throughout the years crime related to attacking person almost remained steady after a negligible decrease in 2003 and 2004 respectively. Breaking and entering of both resident and commercial properties decreased until 2010 and since then there is very little increase. On the other hand, it is noticeable that theft is on the rise since 2011 after a steady decline for almost a decade. To sum up, there is an increase in crime rate in Vancouver since 2011 and it is being contributed primarily by two factors. As we have seen from the previous figure, the crime actually increased in the highly populated and from this figure we also learned that it not the other type of crimes that increased significantly rather it is the theft which increased. Therefore we can deduce that theft in highly populated neighborhoods increased quite a lot since 2011 which is causing overall increase in the crime rate of Vancouver over the last few years.

4.4 Type of crime and its relation to season?

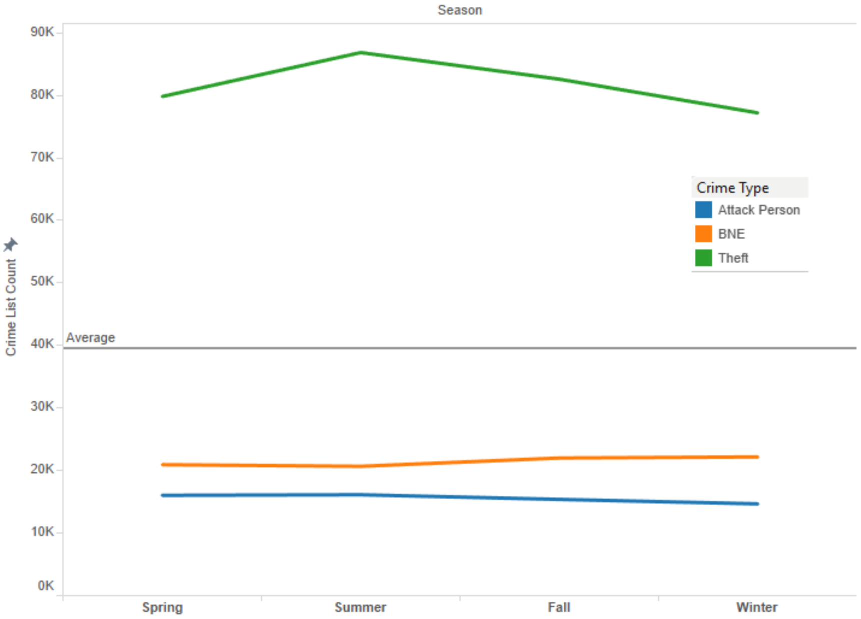


Figure 15: *This figure shows the crime rate for different seasons for all years from 2003 to 2015 in Vancouver.*

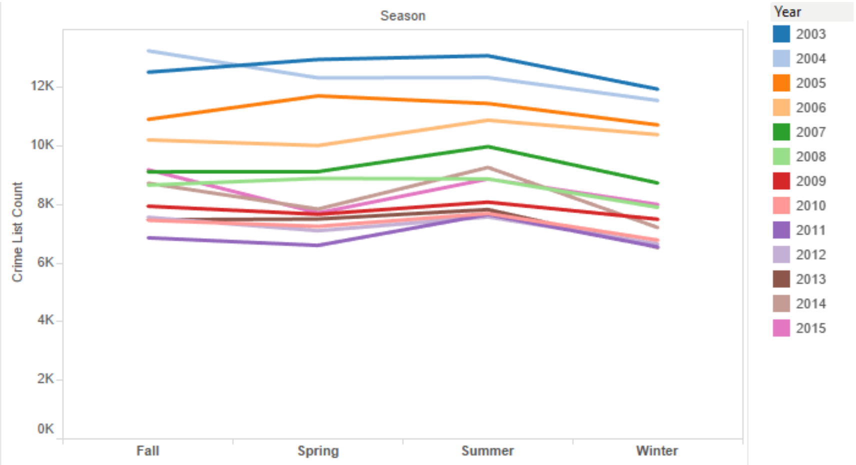


Figure 16: *This figure shows the crime rate for different seasons for each year from 2003 to 2015 in Vancouver.*

We have previously noticed that theft is the most prevalent crime in Vancouver. In this chart which represents the number of different crimes happening in Vancouver during different seasons. Attack on person and Breaking and entering of properties remains almost steady throughout every seasons except in Winter when public attack decreases slighting but breaking and entering increase slightly. One of the reason for this can be that during Winter less people can be found on the street at night as a result attacking on person also decreases. At the same, since people go for holidays during the year end, therefore breaking and entering of properties increases. However, the most noticeable change is the sharp increase of theft during summer time compare to other seasons.

4.5 How much are the rich neighborhoods affected by the crime compared to the others?

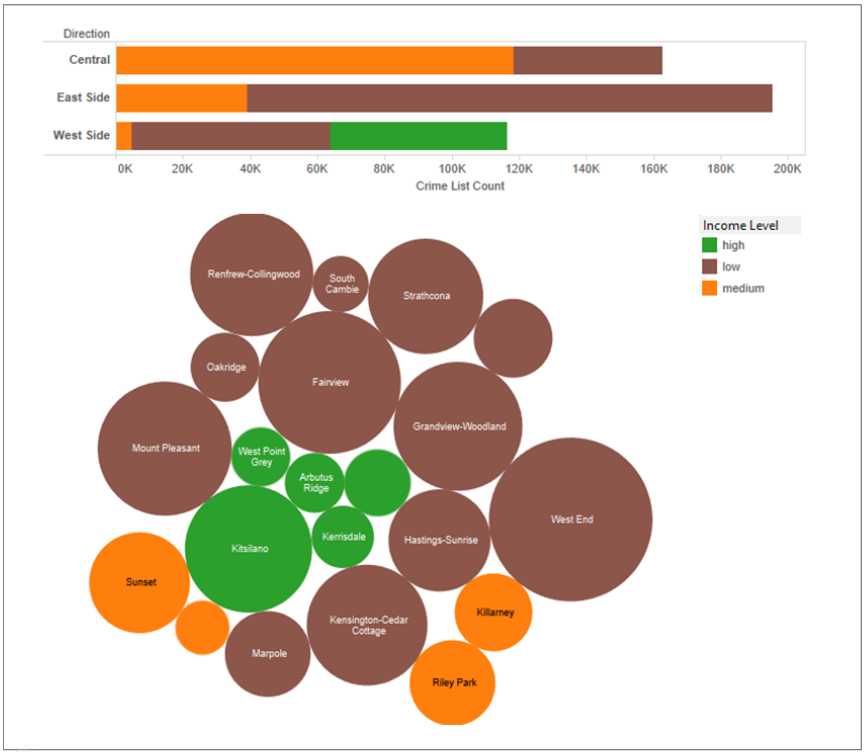


Figure 17: *This figure shows the crime rate for different directions and income level of neighborhoods in Vancouver. The size of the circle represents the crime rate. The bigger the circle, the higher the crime rate.*

The least crime affected areas in Vancouver are the neighborhoods where income level (median) is high (over CAD 90,000). From the figure above we can notice that the richest neighborhoods in Vancouver are all in the west side. It is also noticeable that the west side has comparatively less crime happening than the other sides of Vancouver. To understand whether crime is correlated with the income level let us look at the crime bubble shown just below the bar charts in the figure above. The size of the circle represents the number of crimes in that particular neighborhood. This gives us a clear view that neighborhoods where the median income level is low, represented with the color brown, has higher crime rates since the circles for these neighborhoods are comparatively larger than that of neighborhoods where median income level is medium or higher. Kitsilano is the only exception in this circumstance, where the crime rate is higher despite it being one of the rich neighborhood.

4.6 Downtown, the crime hotspot of Vancouver?

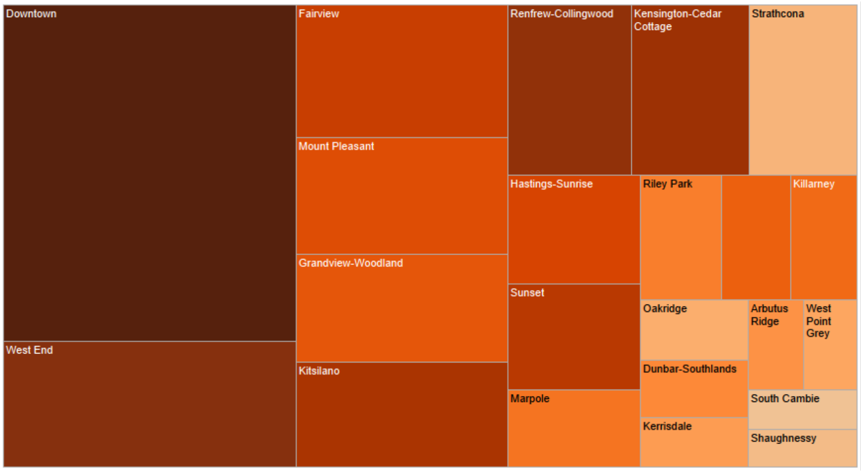


Figure 18: *This figure shows the crime rate for all neighborhoods in Vancouver. The size of the square represents the crime rate and the color represents the population. The bigger the square, the higher the crime rate. The deeper the color, the larger the population.*

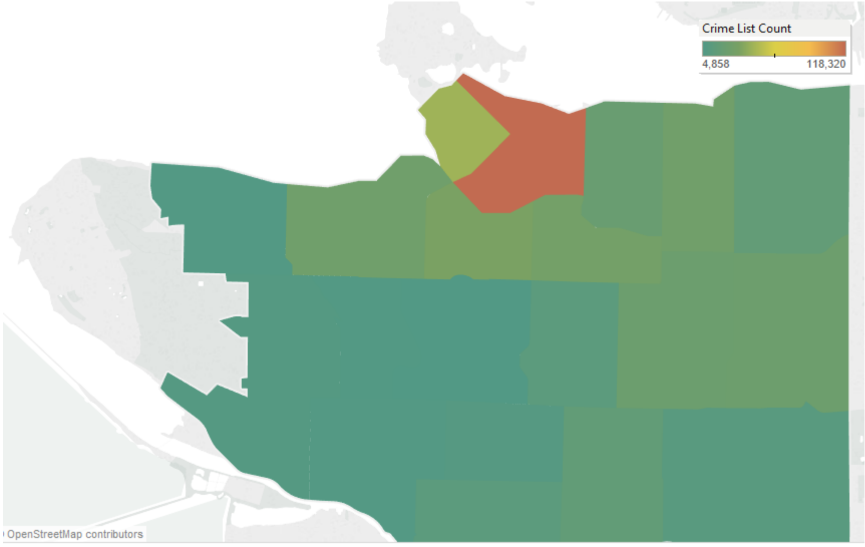


Figure 19: *This figure shows the crime rate map for all neighborhoods in Vancouver. The warmer the color, the higher the crime rate.*

Downtown is the biggest neighborhood of Vancouver containing its own subareas such as Yaletown, Gastown and Downtown Eastside. In the tree map provided above, each square represents a neighborhood in Vancouver. The size of the square indicates the crime rate in that particular neighborhood and color represents the population respectively. Darker color means more people live in that area and lighter color means less people live in that area. From this we can conclude that Downtown is the most populated neighborhood of Vancouver. At the same time, it is vivid that Downtown has the highest crime rate. The map visualization provided below gives a better intuition. The red area represent Downtown had around 118,320 crimes from 2003 to 2015. We can see that the rest of the Vancouver is labelled with different shades of green indicating a lot less crime happening there compare to Downtown.

5 Conclusion

This project was built with the goal to provide a geo-visualization system which will aid users to answer any queries related to Vancouver crime situation. Data were collected from different sources and was used to clean and build a four dimensional hierarchical cube. The cube was then used to build the geo-visualization system as well as some important insights were found, for instance, theft is increasing in Vancouver since 2011 and its prevalent during the summer time. In the future, the system can be improved more by adding more demographic and socio-economic data. At the same time a data crawler can be designed to retrieve updated data from the online to feed a transactional cube.

References

[1] Vancouver Police Department. Crime heat maps. http://vancouver.ca/police/CrimeMaps/. Accessed: 2016-03-15.

[2] Garner Clancey. Introduction to crime data analysis. http://garnerclancey.com/pdfs/cpp_IntroductiontoCrimeDatasAnalysisPoster.pdf. Accessed: 2016-04-02.

[3] ETHAN ZHE ZHANG. Crime data visualization. http://cargocollective.com/mr_z/Crime-Data-Visualization. Accessed: 2016-04-05.

[4] Vancouver Police Department. Vancouver crime data. http://data.vancouver.ca/datacatalogue/crime-data.htm. Accessed: 2016-03-16.

[5] Vancouver Planning and Development Services. Vancouver census data 2011. http://data.vancouver.ca/datacatalogue/censusLocalAreaProfiles2011.htm. Accessed: 2016-03-16.

[6] Vancouver GIS and CADD Services. Vancouver map data. http://data.vancouver.ca/datacatalogue/localAreaBoundary.htm. Accessed: 2016-03-16.

[7] Oracle. Database data warehousing guide. https://docs.oracle.com/cd/B19306_01/server.102/b14223/ettover.htm. Accessed: 2016-03-16.