Crime Analysis for Neighborhoods in Vancouver

Introduction:

As everyone knows, Vancouver is one of the best place to live in. But which Neighbourhood should you choose and which characters should you consider? Well, most people may say price and safety! Well, there are a lot of websites that can help you to find houses under the budget very easily. So, this project will help you to find how safety that house is.

In details, it will give you Neighbourhood guides in Vancouver based on the crimes reported in each Neighbourhood and clustering those Neighbourhoods into different groups.

Data:

In this report, two datasets will be used. The first one is the crime.csv which includes all crimes reported in Vancouver. It also shows the detail of each crime such as: crime type, date, and Neighborhood. This data is directly download from Kaggle, https://www.kaggle.com/agilesifaka/vancouver-crime-report

The Second dataset is the location data, which combined with a list of Neighborhoods and their GPS location. When I'm doing the research, there' no direct files that contain this information, so I manually create a Location.xlsx file getting the location data from Wikipedia. The purpose of getting this location data is helping to combine the crime data with the Foursquare location data.

Data CleanUp:

Based on the Vancouver Neighborhoods that identified in Location.xlsx dataset, some rows in crime.csv will be deleted (ie. rows that have Neighborhood values: 'Musqueam', 'Stanley Park', or 'NaN').

Also, in Neighborhood column of crime.csv table, 'Central Business District' will be changed to 'Downtown' just for convenience.

After that, pandas' merge function is used to join crime table to Location table based on their Neighborhoods. Once two tables are merged, some columns with useless information are dropped (ie. 'MONTH', 'DAY', 'HOUR', 'MINUTE', 'HUNDRED_BLOCK', 'X', 'Y').

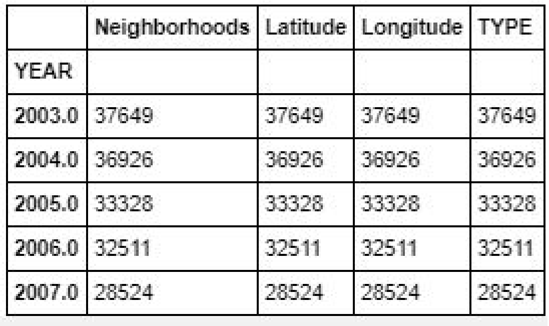
Methodology:

1. Year vs. Crimes

Based on the purpose of this report, the first things I want to do is finding the trend of total crimes number in Vancouver. In other words, is the crime rate increased or decreased in the past few years? To do that, the Vancouver's crime data are grouped by the happened time, and then summed up to get the total Crimes number in each year. After that, the grouped data is sorted by its year. bar chart is used to represent total crimes number every year from 2003 to 2019.

2. Neighborhoods vs. Crimes

After finding out when the crime happened, I want to know where it happened. In crime table, there is a column called Neighborhoods, which is a neighborhood boundaries that Vancouver used break up city's geographic area for delivering services and resources.

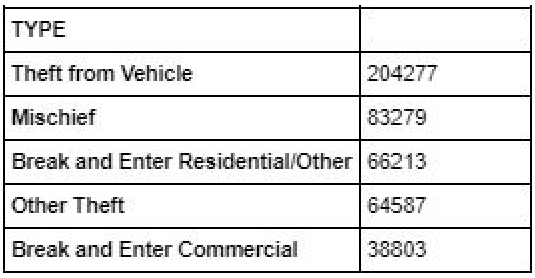


	Neighborhoods	Latitude	Longitude	TYPE
YEAR				
2003.0	37649	37649	37649	37649
2004.0	36926	36926	36926	36926
2005.0	33328	33328	33328	33328
2006.0	32511	32511	32511	32511
2007.0	28524	28524	28524	28524

The crimes data is grouped by their Neighborhoods and stored in a new dataframe. Then it is sorted by the total crimes number in each Neighborhood and a bar chart is plotted for better understanding.

3. Type of Crimes vs. Crime rates

Again, the crimes data is grouped by the it's type in order to find the crime type that happened most often in Vancouver. The top five common crime types are:

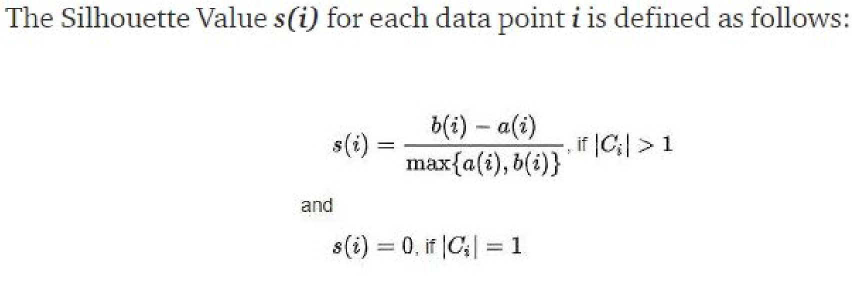


TYPE	
Theft from Vehicle	204277
Mischief	83279
Break and Enter Residential/Other	66213
Other Theft	64587
Break and Enter Commercial	38803

4. Neighborhoods Clustering and Mapping

On of the most important things to do is clustering all Vancouver Neighborhoods into different groups. To do that, the python built-in function 'KMeans Clustering' is used to clusters Vancouver Neighborhoods into different groups based on the Venus categories in each Neighborhood. But what cluster numbers should I use? Or, how many clusters should it be?

“The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters.” (Wikipedia).



The Silhouette Value $s(i)$ for each data point i is defined as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_i| > 1$$

and

$$s(i) = 0, \text{ if } |C_i| = 1$$

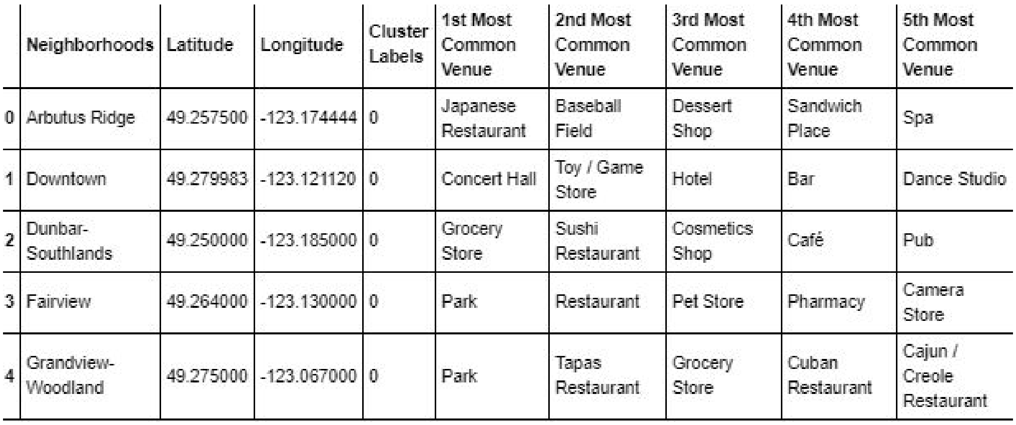
Source: Wikipedia

Well, the function called ‘The Silhouette Method’ is used and a line plot is present for convenience. By looking at the local maximum of the result plot, it seems that cluster number 3 is good to use.

For the factors used to clustering crime data, I’m using ‘Foursquare’ to request information about each Neighborhood in Vancouver. More specifically, getting all the venues in the Neighborhood with details of each venue like their name, category and location. The data gathered from ‘Foursquare’ includes 85 unique venue categories in total.

```
In [23]: print('There are {} uniques categories.'.format(len(van_venues['Venue Category'].unique())))
         There are 85 uniques categories.
```

Now, we can use K-means Clustering to separate Crimes into 3 clusters, which are labeled as 0, 1, 2. After merge it into dataframe as Column ‘Cluster Labels’, it looks like this:



	Neighborhoods	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Arbutus Ridge	49.257500	-123.174444	0	Japanese Restaurant	Baseball Field	Dessert Shop	Sandwich Place	Spa
1	Downtown	49.279983	-123.121120	0	Concert Hall	Toy / Game Store	Hotel	Bar	Dance Studio
2	Dunbar-Southlands	49.250000	-123.185000	0	Grocery Store	Sushi Restaurant	Cosmetics Shop	Café	Pub
3	Fairview	49.264000	-123.130000	0	Park	Restaurant	Pet Store	Pharmacy	Camera Store
4	Grandview-Woodland	49.275000	-123.067000	0	Park	Tapas Restaurant	Grocery Store	Cuban Restaurant	Cajun / Creole Restaurant

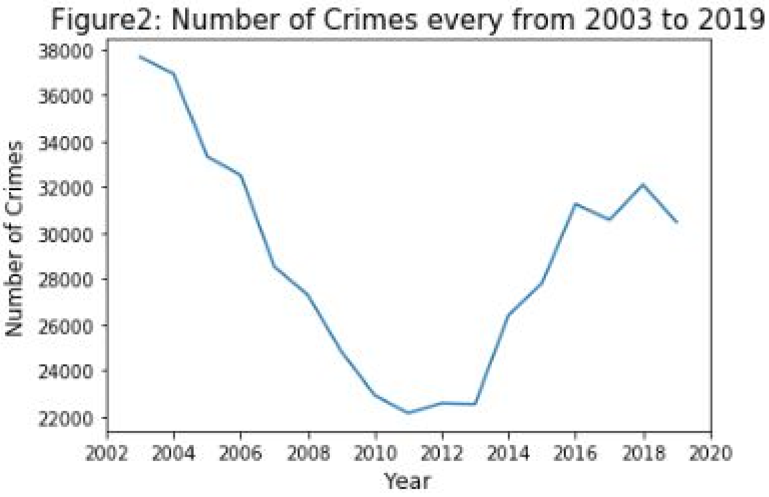
After clustering Vancouver Neighborhoods into three groups, I want find out the percentage of the most common crime types in each cluster. I'm using the word plot and pie chart to present the frequency of crimes in each cluster.

Lastly, in order to show the clusters in better visualization, a Folium map is plotted to represent Neighborhoods within different clusters as different color. As shown in the Results section.

Results:

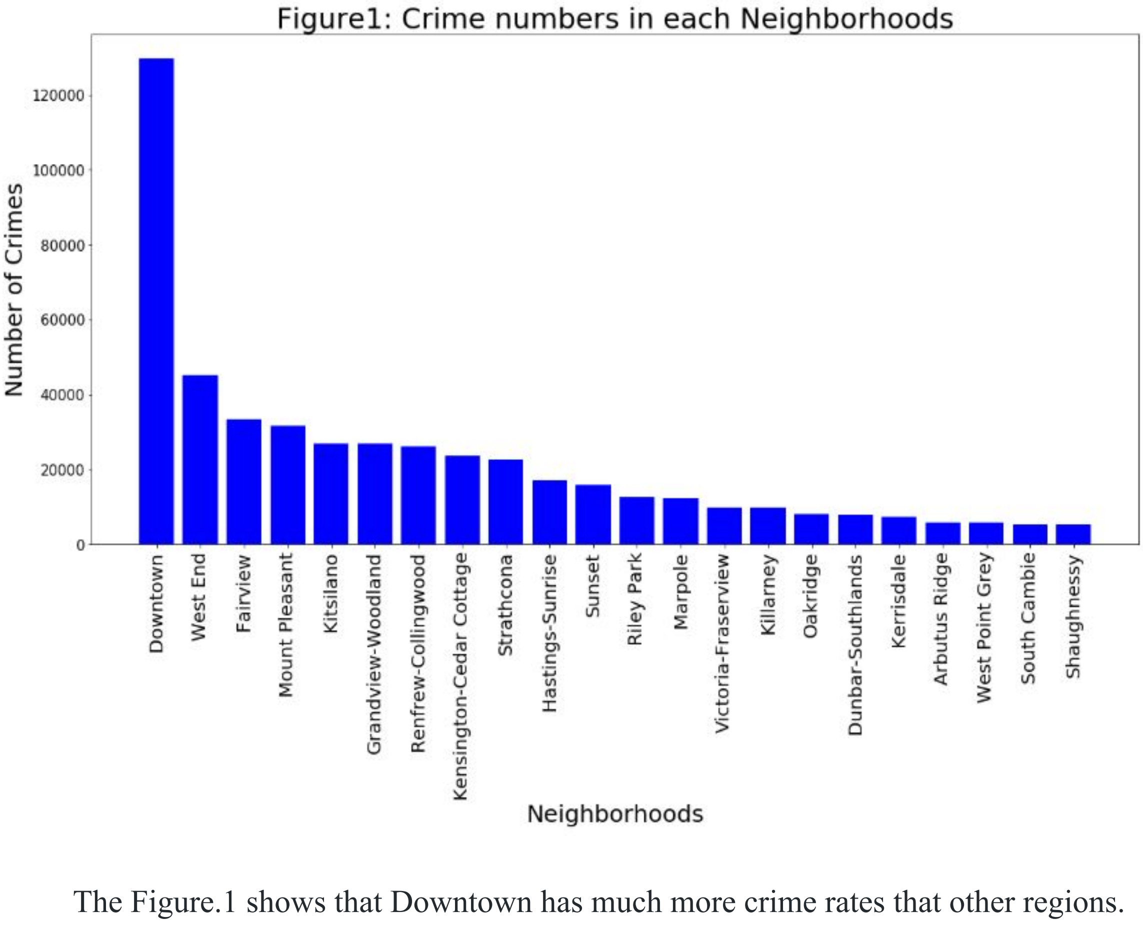
1. Year vs. Crimes

By looking at total crimes rates every year in Figure.2, we can find that the crime rate reaches a local maximum in 2018, and local minimum from 2010 to 2013. Even though the crimes rates starts increase since 2012, the crime number per year is still much more lower than the number in year 2003. Also, the crime rate shows a decrease trend since 2018. So, we can say that Vancouver becomes safer since 2003.



2. Neighborhoods vs. Crimes

There are 22 unique Neighborhoods in Vancouver are used in this report.

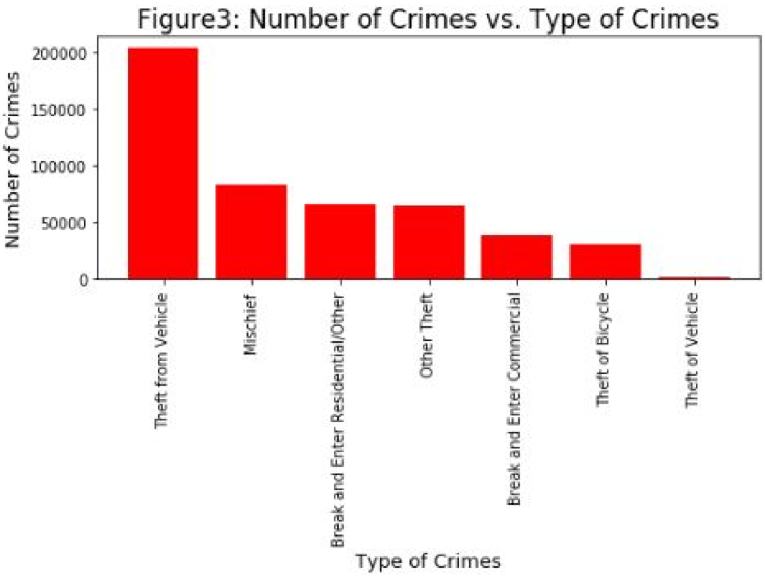


The Figure.1 shows that Downtown has much more crime rates that other regions. And the five most dangerous regions are 'Downtown', 'West End', 'Fairview', 'Mount

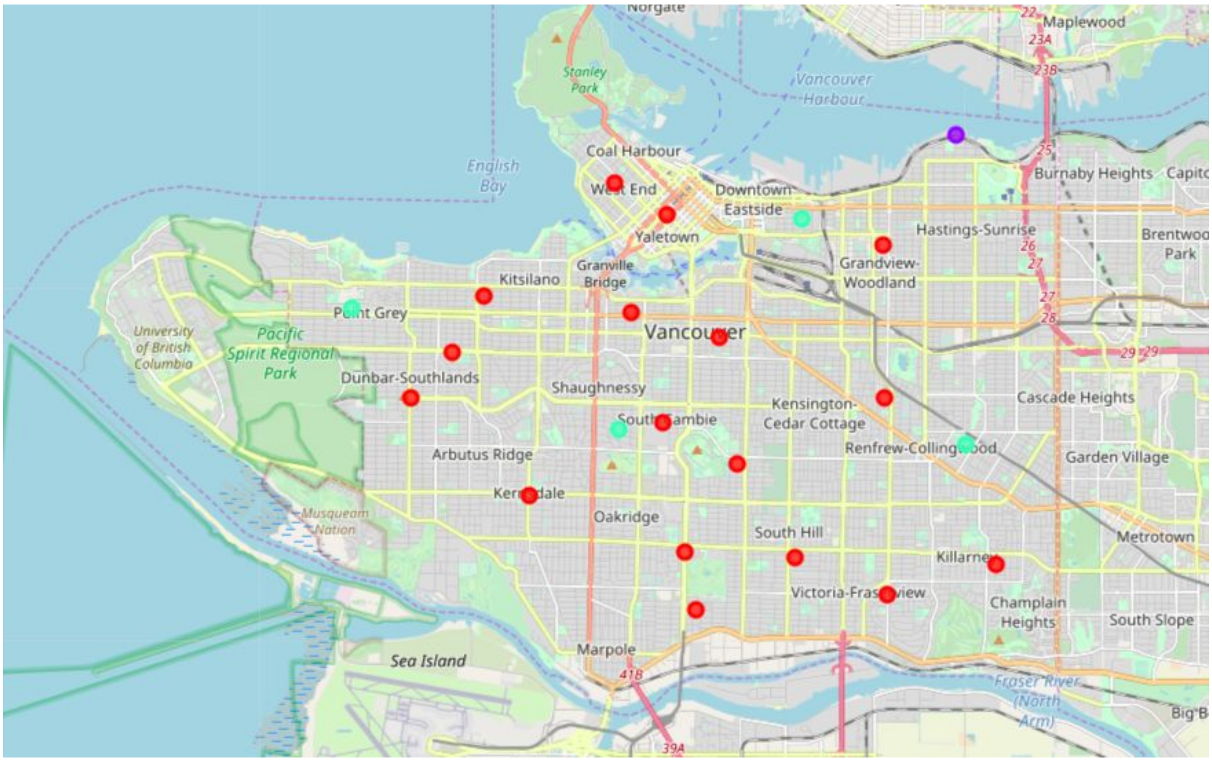
Pleasant', 'Kitsilano'. From these findings, we can say that most crimes are happen at the top part of Vancouver.

3. Type of Crimes vs. Crime rates

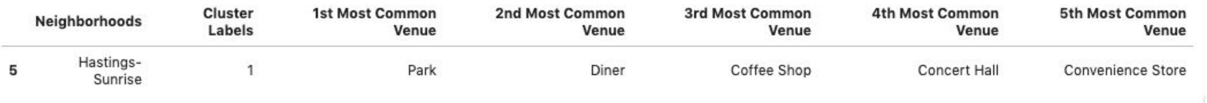
After grouping crime data by its types, Figure.3 shows that the top five crime types in Vancouver are: 'Theft from Vehicle', 'Mischief', 'Break and Enter Residential/Other', 'Other Theft' and 'Break and Enter Commercial'. So, we can say that most crimes in Vancouver are Financial related.



4. Neighborhoods Clustering and Mapping



The folium map is shown above. It can be easily seen that Neighborhood ‘Hastings-Sunrise’ is very different from all others as it’s top five venue categories are ‘Park’, ‘Diner’, ‘Coffee Shop’, ‘Concert Hall’ and ‘Convenience Store’, which seems a good place to live.



	Neighborhoods	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
5	Hastings-Sunrise	1	Park	Diner	Coffee Shop	Concert Hall	Convenience Store

In more details, the word plot and pie chart are plotted for each clusters as shown below:

In cluster one, there are 42 percent ‘Theft from Vehicle’, 16 percent ‘Mischief’, and 14 percent ‘Other Theft’.

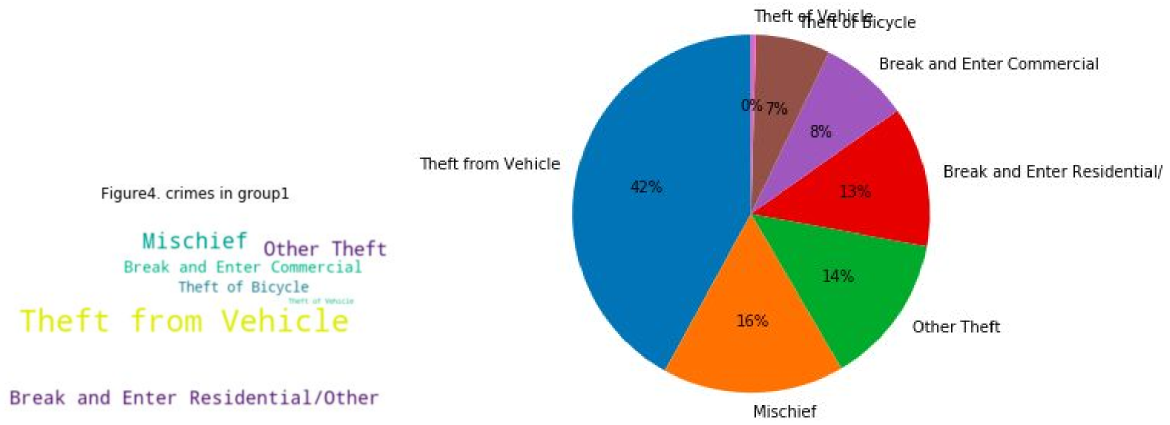


Figure4. crimes in group1

In cluster two, there are 41 percent ‘Theft from Vehicle’, 20 percent ‘Break and Enter Residential’ and 20 percent ‘Mischief’.

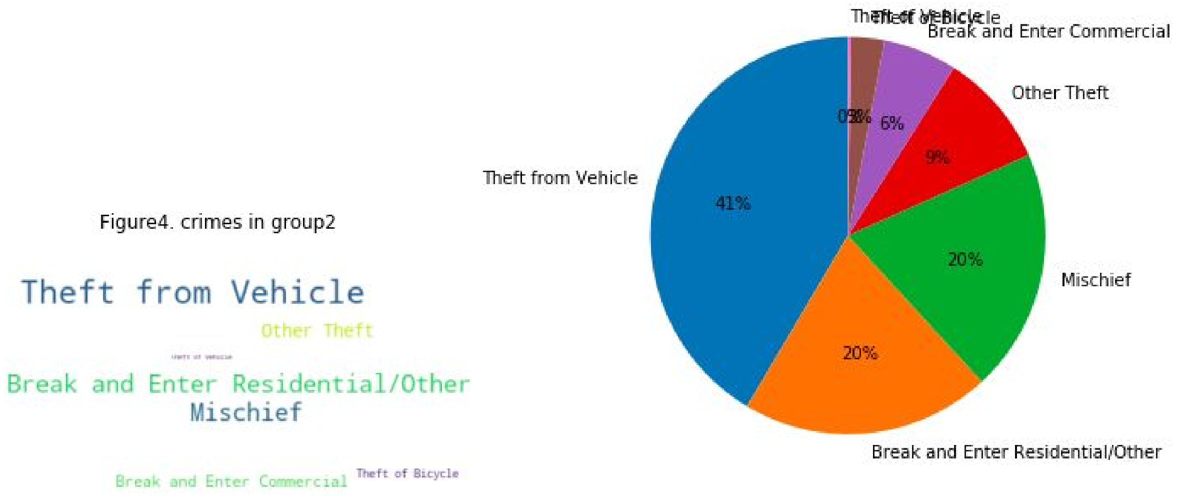


Figure4. crimes in group2

In cluster three, there are 40 percent ‘Theft from Vehicle’, 20 percent ‘Mischief’, and 17 percent ‘Break and Enter Residential’.

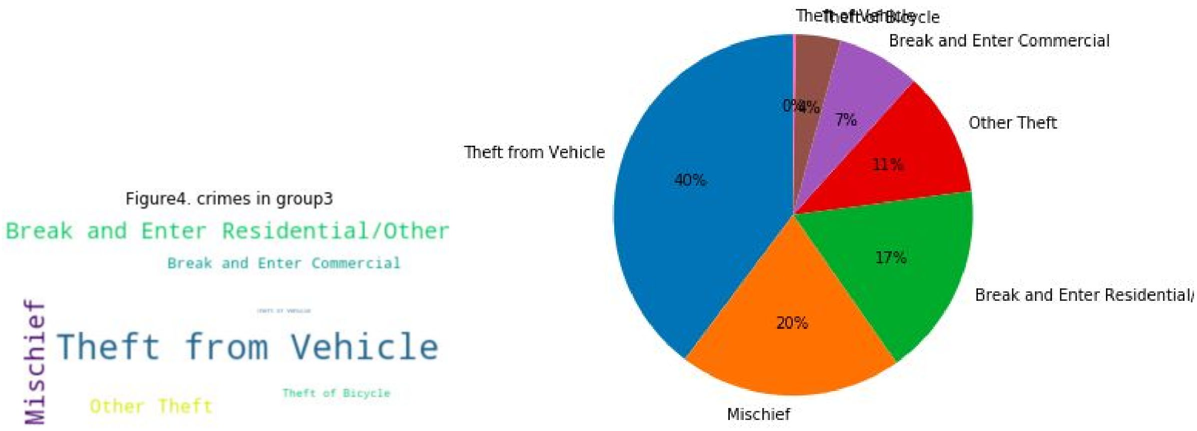


Figure4. crimes in group3

By looking at all three cluster's crime data, most crimes in Cluster one are theft related, adn only 8 percent crimes are 'Break and Enter Residence'. In Cluster 2 and 3, even though most of crimes are still Theft related, we can see that 'Break and Enter Residence' is also a serious issue as it's about 20 percent in these clusters. As people usually concern Residencial related crimes when they are choosing places to live, we can conclude that Neighborhoods belongs to Cluster 2 and 3 are less safer than other Neighborhoods in Vancouver.

Discussion:

In this report, I'm using the method K-means Clustering to cluster the crime data. There are some other Clustering methods can be used such as Hierarchical Clustering and Density-Based Clustering. Further study can try different method and check the error to find the best option.

Also, this report used venues which got from FourSquare API to identify Neighborhoods into different group. This maybe not best way to cluster them. Some other data like average house price or population in Neighborhoods can be used as factors to cluster Vancouver Neighborhoods.

Finally, the type of crimes in the dataframe can be more specific. If we can find a dataset with more detailed information, then there may have more similarities and differences found.

Conclusion:

Based on all the findings shown in this report, we can conclude that Vancouver becomes safer since 2003. By analyzing data, it shows that 'Downtown' has the most crime rates in Vancouver. But don't worry too much because most of crime in Vancouver are Theft related like 'Theft from Vehicle'. Neighborhoods with high 'Break/Enter Residencial' rate are less secure for people to live, which are ‘Hastings-Sunrise’, ‘Renfrew-Collingwood’,’Shaughnessy’, ‘Strathcona’ and ‘West Point Grey’.