**ETL Project**

**Crime comparison between**

**Toronto and Vancouver**

Joyce Jian [yqjianyq@gmail.com](mailto:yqjianyq@gmail.com)

Sherry Huang [hyhuan0318@gmail.com](mailto:hyhuan0318@gmail.com)

Sherry Huang [yidan.h77@gmail.com](mailto:yidan.h77@gmail.com)

Ruozhuo Wang [zara.wang85@gmail.com](mailto:zara.wang85@gmail.com)

**2019 August**

**1.0 Extract Data**

**1.1 Data sources**

In this project, two original criminal datasets were retrieved and downloaded as CSV files.

* Criminal data from Toronto Police:

MCI Table: <https://data.vancouver.ca/datacatalogue/crime-data-details.htm>

Homicide Table: <http://data.torontopolice.on.ca/datasets/7826a3ef2eff4d64a7f70e909de007b5_0>

Traffic Table: <https://data.torontopolice.on.ca/datasets/ksi>

* Criminal data from City of Vancouver official website: <https://data.torontopolice.on.ca/pages/catalogue>

**1.2 Data columns**

The crime rate in Vancouver dataset (2003-2018) contains the following information:  
- Crime type  
- Crime date  
- Crime time  
- Neighbourhood

The crime rate in Toronto MCI Table (2014-2018) contains the following information:  
- Offence type  
- Reported month  
- Reported day of the week  
- Occurrence month  
- Occurrence date of the week  
- Longitude  
- Latitude

**2.0 Transform**

**2.1 Questions**

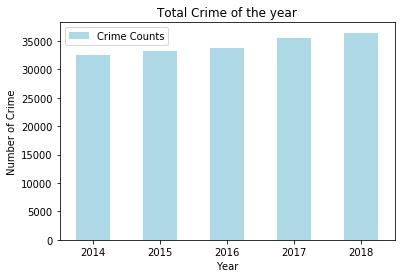
We try to answer the following questions based on the original datasets:

* Is there a year-over-year change on the total number of crimes in each of the two cities?
* Is there a month-over-month change on the average number of crimes in each of the two cities?
* Which type of crime dominates in number?
* Do crimes usually happen during the day (6:00-18:00) or night (18:00-6:00)?

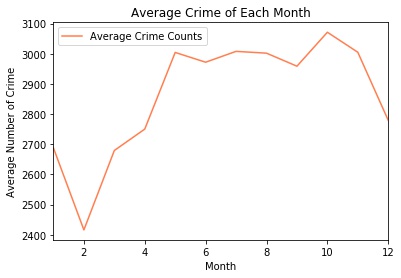
**2.2 Data cleaning and analyzing**

**2.2.1 Crime data of the City of Toronto**

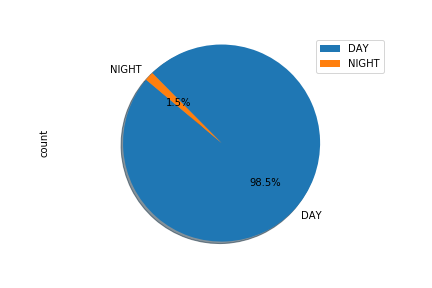
* Imported dependencies and read in csv datafiles
* For all three tables:
  + Extracted and stored Year and Month as new columns from Occurrence Date column
  + Copied necessary columns only to a new dataframe
  + Set up new column to category occurence month as Day or Night
  + Modified columns data types and sorted the columns order so that they would conform with each other
* For MCI Table: Change Crime Type “Theft Over” to “Other Theft” to confirm with data in Vancouver dataset
* For Homicide and Traffic Tables: Added an extra column stating the crime type
* Concatenated all three tables; renamed the columns and modified the ID to starts from 1
* The trend of yearly total crime counts from 2014 to 2018.
  + Used “groupby” and “count” to get the total crime count of each year



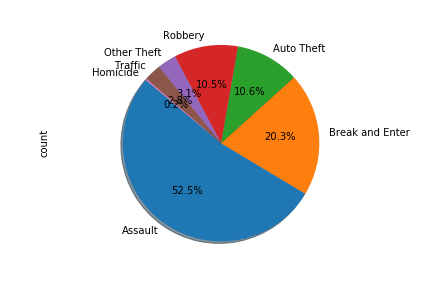
* Crime counts by months of the year
  + Grouped the data by “month” and “year” and counted the number of crimes.
  + Grouped the result of the above by “month” and got the average count.



* Total crime counts by “time” of a day
  + Used “groupby” and “count” to get the total crime count that happen day and night.

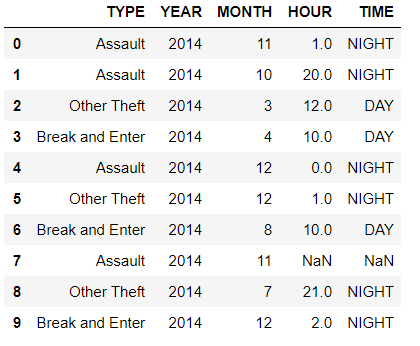


* Total crime counts by “type”
  + Used “groupby” and “count” to get the total count of each crime type

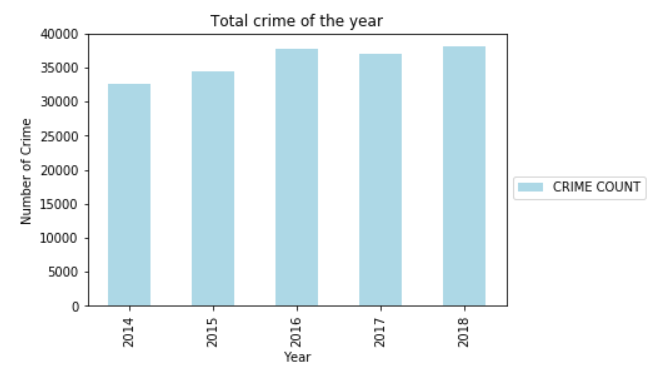


**2.2.1 Crime data of the City of Vancouver**

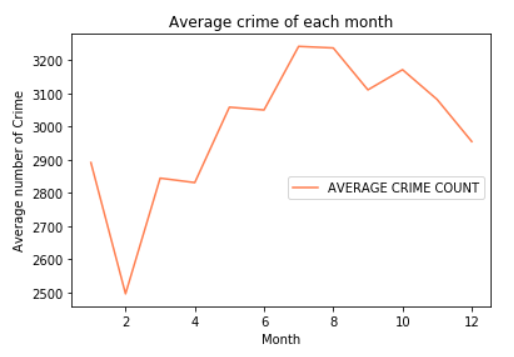
* Imported dependencies.
* Used “pd.read\_csv” to read the csv data file as DataFrame.
* Created a new DataFrame that only includes the columns will be used.
* In the column of “type”, changed the crime category to be compliant with the data of the City of Toronto. After adjustment, there are 6 types of crime (Other Theft, Assault, Break and Enter, Traffic, Auto Theft, and Homicide) in the data of both cities.
* In the column of “hour”, replaced the missing data with “NaN”.
* Inserted a new column named “time” with information “night”, “day” or “NaN” depending on the column “hour”. The crime happens during the time period from 6 (included) to 18 (not included) was classified as “night”, otherwise, it was classified as “day” or “NaN”.
* Showed the cleaned data.



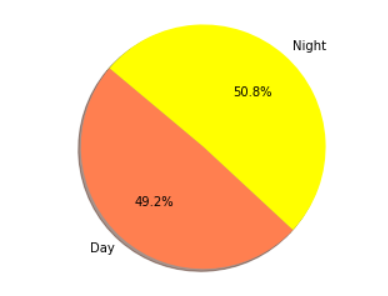
* The trend of yearly total crime counts from 2014 to 2018.
  + Used “groupby” and “count” to get the total crime count of each year.



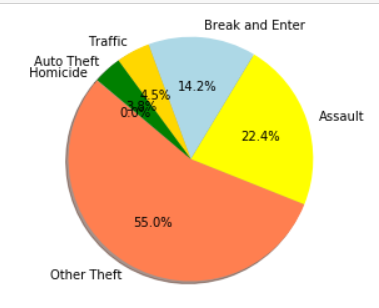
* Crime counts by months of the year
  + Grouped the data by “month” and “year” and counted the number of crimes.
  + Grouped the result of the above by “month” and got the average count.



* Total crime counts by “time” of a day
  + Used “groupby” and “count” to get the total crime count that happen day and night.



* Total crime counts by “type” of the crime
  + Used “groupby” and “count” to get the total count of each crime type

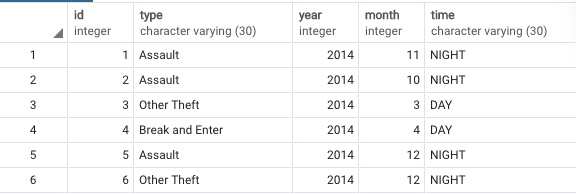


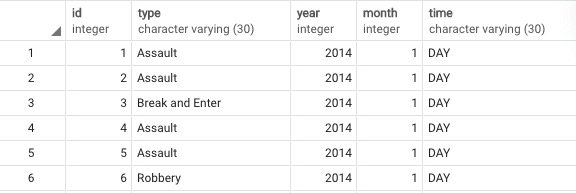
**3.0 Load:** the final database, tables/collections, and why this was chosen.

* In SQL, created a database called “crime\_db” and create 2 tables called “vancouver\_crime” and “toronto\_crime” respectively.
* Used “engine” to connect to the local database and check for tables.
* Used pandas to load cleaned data into database “Crime\_db”.
* Confirm data has been added by querying tables “vancouver\_crime” and “toronto\_crime”



* In SQL, use “select \* from …” to check the uploaded data of each table.





* Used SQL queries to select, groupby(types/years) and count criminal data from both raw tables as desired into temporary tables.
* Joined temporary tables with grouped data and generated a new table with criminal counts comparison between two cities by years and types.

