# 3250 Group Project: Data cleansing process

Our team leveraged the X data-set that focused on robberies in the Toronto area (“Robbery” data) occurring between the years [**2014** and **2018**](http://data.torontopolice.on.ca/datasets/d888b0ed52864bd6ace51888322fda58_0).

Overall, the “Robbery” data-set was largely clean. There were **29** columns and **18128** rows. **100**% of the rows had every cell populated. [Figure **1**](#_Artifacts)shows the columnar information for the source data. There were **13** obvious categorical opportunities and through the data engineering process a further **20** categories were included (*See summary below*).

There were opportunities to integrate the “Robbery” data with additional third party database tables. Refer to the attached data-sets for the associating columnar details:

* A robust **Dates** data-set which not only included calendar dates information, but also the *Part of Week*, *Time of Month* and also the *Holiday* information for Ontario. This data can be applied to the Occurrence and Reported dates fields.
* The Long & Lat fields can reveal **Postal Codes** based geometrical calculation between the central longitude and latitude of the Postal Code.
* The Occurrence and Reported hour columns allowed for the determination of the **Time of Day** data-set. Which also can determine the Sky Light Category.

This gave rise to normalization approaches as discussed with Saad. SQL was used to achieve this normalization. Refer to attached SQL Script used.

In *summary*,

* OOTB Features: ***offence, reportedyear, reportedmonth, reportedday, reporteddayofyear, reporteddayofweek, occurrenceyear, occurrencemonth, occurrenceday, occurrencedayofyear, occurrencedayofweek, neighbourhood, premisetype***
* Engineered features: ***Occurrence\_TimeOfDayCategory, Occurrence\_SkyLightCategory, Reported\_TimeOfDayCategory, Reported\_SkyLightCategory, Occurred\_PostalCode, Occurred\_FSA, PopulationCount, PopulationDensity, PopulationCount\_AgeCohort, Occurred\_TimeOfWeek, Occurred\_TimeOfMonth, Occurred\_IsHoliday, Occurred\_CalendarQuarter, Reported\_TimeOfWeek, Reported\_TimeOfMonth, Reported\_IsHoliday, Reported\_CalendarQuarter*** ***, Reported\_After\_Occurence\_WEEKS, Reported\_After\_Occurence\_DAYS, Reported\_After\_Occurence\_HOURS***

# Artifacts

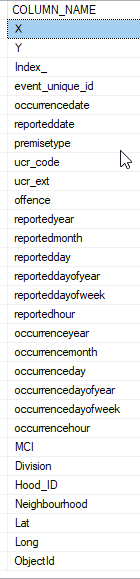


Figure 1: Original Data-set