# 3250 Group Project: Data Engineering Journey

This section of the project focuses on the steps taken to prepare the raw data into a consumable format within a Jupyter Notebook (“Notebook”) for data analysis by some of members of the team. Additionally, within this section includes, the Notebook augmentation process following the completion of the various data analysis.

## Data Sourcing

After considering different data-sets from Women’s rights to Housing, the team settled on leveraging the Toronto Police Robbery data-set[[1]](#footnote-1) (“Robbery”) that spanned the years 2014 through to 2018. This data is publicly available [here](http://data.torontopolice.on.ca/datasets/d888b0ed52864bd6ace51888322fda58_0) and was downloaded in a CSV format.

The “Robbery” data-set contained twenty-nine (29) different columns and contained over eighteen thousand (18,128) records about different robberies that were reported between 2014 and 2018 inclusively. [Figure **1**](#_Artifacts_1)shows the columnar information for the source data.

At a glance, there are several useful data-points for analysis such as the date and time when the incident occurred against when it was reported, the type of robbery incident, the specific GPS coordinates where it occurred, the Police division that recorded the incident and the edifice structure which was burgled.

## Data Wrangling

In the course, we were trained to use Python to manipulate data; but were encouraged to use other tools if they were more effective at performing certain aspects of data manipulation or data visualization than Python. From experience with data-sets similar in structure to the “Robbery” data-set, it was determined that there were some normalization opportunities – and as such SQL was employed to manipulate the data into different sub data-sets. Specifically, within SQL Server 2017. The SQL script ([robberies.sql](#_Artifacts_1)) that was used to manipulate the data is attached as an artifact to this section.

### Data Feature Assessment

Overall, there were thirteen (13) obvious categorical opportunities and through the data engineering process, a further twenty (20) categories were included:

* Out of the box (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\_TimeOfMonth, Occurred\_IsHoliday, Occurred\_CalendarQuarter, Reported\_TimeOfWeek, Reported\_TimeOfMonth, Reported\_IsHoliday, Reported\_CalendarQuarter* *, Reported\_After\_Occurrence\_WEEKS, Reported\_After\_Occurrence\_DAYS, Reported\_After\_Occurrence\_HOURS, Reported\_After\_Occurrence\_MINUTES*

### Data Augmentation

There were opportunities to integrate the “Robbery” data with additional publicly available Dates & Forward Sortation Area (FSA)[[2]](#footnote-2) database tables. Refer to the attached [SQL script](#_Artifacts) for the detailing the procedure to connect to those data-sets:

* Dates\*
* FSA
* Neighbourhood\*

\* These are explained in the next sub-section.

### Data Normalization

The data-normalization process broke the whole “Robbery” dataset into 6 subsets.

1. Robbery data 🡺 this is the main fact data
2. Offences 🡺 a key/ value pairing cataloging the types of offenses
3. Dates\* 🡺 a data-set about different aspects about a day.
4. Neighbourhood\*\* 🡺 a data-set about different aspects of a Toronto neighbourhood
5. Premise Types 🡺 a key/ value pairing cataloging the types of edifices
6. Time of Day\*\*\* 🡺 a data-set about different aspects of the time of day

\* Dates: This is a robust data-set which we authored leveraging a SQL Server stored procedure. It not only included calendar dates information, but also the *Part of Week*, *Time of Month* and we’ve manually layered in the *Holiday information for Ontario*. This data can be applied to the Occurrence and Reported dates fields to further understand questions like when in the month do most robberies take place? Or do most robberies take place on weekends or weekdays? Or what is the most common day of the week, month, year for robberies?

\*\* Neighbourhood: This census-driven, publicly available data-set was manipulated to provide additional insights into Toronto neighbourhoods. This data was made available [here](https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/community-services/#8c732154-5012-9afe-d0cd-ba3ffc813d5a).[[3]](#footnote-3)

\*\*\* Time of Day: The “Robbery” data-set included **Occurrence hour** and **Reported hour** columns which allowed for the determination the Sky Light Category – i.e. did the robbery occur in daylight or at night? Was it reported in the day or at night?

### Data Purging

Overall, the “Robbery” data-set was largely clean. **100**% of the rows had every cell populated. However, upon a more in-depth inspection, we determined that it would be better to drop all the records that related to incidents that occurred prior to Jan 1, 2014 (*These records were a part of this data-set as they were reported between the years 2014 and 2018*). This action removed 43 records to yield a 18085 record “Robbery” data-set.

## Data Delivery

The six (6) aforementioned data-sets are packaged within a “Notebook” and uploaded to Github.com for consumption by the team.

Setting up GitHub account:

GitHub is the largely used code collaboration platform worldwide and so we setup the GitHub account, configure the collaborators. The initial notebook file and the clean dataset (csv files) were uploaded to GitHub account found [here](https://github.com/SandipRudani1988/TorontoRobbery) for everyone to clone and do the work separately on their own branch and merge the code via pull requests but we found it confusing and needed the learning curve so we end up merging the code manually.

## Post-Analysis “Notebook” Augmentation

* Make sure that everyone is using the same version of dataset
* The code written to derive each answer works with the final dataset
* Getting rid of any redundant code, prepare common function etc.

## Artifacts

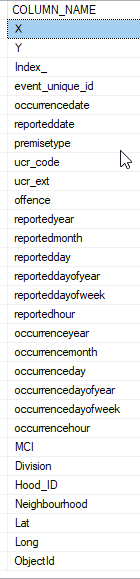


Figure 1 Columnar information about "Robbery" data-set



Figure 2 Code used to manipulate the data

1. <http://data.torontopolice.on.ca/datasets/d888b0ed52864bd6ace51888322fda58_0> [↑](#footnote-ref-1)
2. Theatre location information from Cineplex Entertainment LP throughout Toronto is publicly available. [↑](#footnote-ref-2)
3. <https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/community-services/#8c732154-5012-9afe-d0cd-ba3ffc813d5a> [↑](#footnote-ref-3)