Data Processing

We are using the [Police cases dataset](https://opendata.utah.gov/Public-Safety/SALT-LAKE-CITY-POLICE-CASES-2016/trgz-4r9d) from 2008-2016 from Utah Open Data catalog. Each record in the data corresponds to an instance of reported crime at one of the police stations in Salt Lake City. It contains information about the type, time and date of occurrence and reporting, and location of the crime. There are about 50K+ records per year. The data is open access.

Each year dataset had the following information: Case Number, UCR Code or NCIC Code, Description, IBR Code, Occurred Date, Reported Date, Day of Week, Location, City, City Council, Police Zone, Police Grid, X-Coordinate, Y-Coordinate.

The two main problems with these datasets are:

1. The schema is not the same in each year dataset. For example, some years display the UCR Code for each record while some display the NCIC Code. The column names are also different for every year.
2. Some crime types are not recorded over all the years.
3. The size of each dataset is around 10 MB and trying to load all the data will make the visualization slow.

In order to overcome these problems, we took the following measures:

1. We decided on a schema that would be uniform in data across all the years.
2. We removed all rows in which data from any of the columns in the schema are missing.
3. We considered only a subset of the crimes to visualize which are more significant in the dataset and which are present over all the years from 2008 to 2016.
4. We also grouped certain crimes which were similar to each other.

We processed each year dataset and create a CSV file named ‘[year]\_processed\_nowhitespace.csv’ with the following columns:

**DESCRIPTION:** It is a description of the crime in a particular record. The value is one of the following categories: Assault, Burglary/ Larceny/ Robbery, Damaged Property, Drugs, Homicide, Kidnap, Traffic, Weapons.

The original data had a column for Occurred Date and Reported Date. In most cases the Occurred Date and Reported Date are the same. We considered Occurred Date for finding the valued of **DAY**, **MONTH**, **YEAR** and **TOD** columns for all of the years except 2011, where we considered the Reported Date since Occurred Date was not recorded.

**DAY:** It is the day of the month that the crime in the record was committed.

**MONTH:** It is the month of the year that the crime in the record was committed.

**YEAR:** It is the year that the crime in the record was committed.

**TOD:** It is the time of day that the crime in the record was committed. The 24 hours in a day are divided into 8 parts: 12 am to 3 am, 3 am to 6 am, 6 am to 9 am, 9 am to 12 pm, 12 pm to 3 pm, 3 pm to 6 pm, 6 pm to 9 pm, and 9 pm to 12 am. The value of **TOD** is between 1 and 8 according to which octant it belongs to.

**DOW:** It is the day of the week that the crime in the record was committed. This data was directly taken from the original dataset. It is a value between 1 and 7 depending on the day of the week where 1 means the crime was committed on a Sunday.

**ADDRESS:** It is the street address of the location where the crime in the record was committed. This was also directly taken from the original dataset with very little processing like removing of whitespaces.

In order to populate the Map View, we need the **LATITUDE** and **LONGITUDE**. The X-coordinate and Y-coordinate in the original dataset are State Plane Coordinates and we could not figure out a way to convert them into the corresponding latitude and longitude. Hence, we had to use the street address to find the latitude and longitude. We aggregated all the year datasets and got a list of the unique addresses. Then we tried to use a python package called geopy to find the latitude and longitude for each address. When this method could not find the latitude and longitude of over 3000 addresses out of 9224 addresses, we turned to Google Maps to obtain the latitude and longitudes.

**LATITUDE:** The latitude of the location where the crime in the record was committed.

**LONGITUDE:** The longitude of the location where the crime in the record was committed.

These steps reduced the size of the data files drastically from around 10 MB to 2 MB, but we faced trouble again while implementing the statistics view. The main problem was that we had to process the data again to be able to create Line Charts for the Yearly Statistics, Monthly Statistics, Weekly Statistics, and Hourly Statistics from it. If we did the processing in JavaScript, it was taking a lot of time to process and load the data. Hence, we decide to create separate json files called ‘all\_years.json’, ‘all\_months.json’, ‘all\_weeks.json’ and ‘all\_hours.json’. ‘all\_years.json’ contains the aggregated data over the years for each crime type. ‘all\_months.json’, ‘all\_weeks.json’ and ‘all\_hours.json’ contain the aggregated data over the years for each month of year, day of week or time of day for each crime type.