Project 1 Intermediate Report

1. Project Details

# Project Topic：**Urban Crime Rate Prediction**

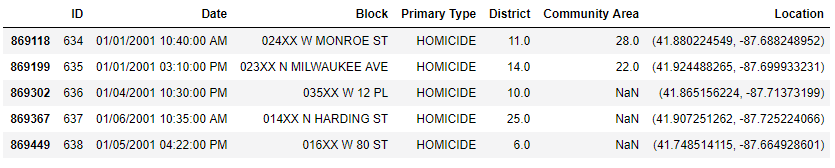
# Team Members: Hongyu Yan, Chenshu Xu

1. Project Goal

The goal of the whole project is to develop a effective predictive modeling framework with the historical crimes data, weather data, traffic data and demographic data. The predictive modeling will be used to predict the possibility that a crime will occur with a specific location in Chicago.

Current goal, at the first phase is to collect the data that we need. Then preprocess the data sets into a new dataframe with the useful datas. At the end, merge all sub datasets into a final database which will be used to calculate the crime rate, build the Model and doing the model evaluation.

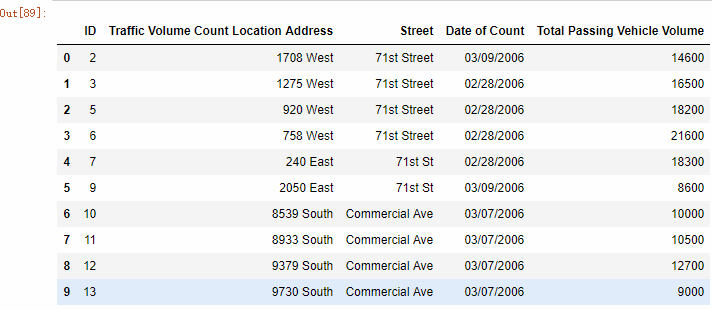
1. Data Sets
   * Crime data set which contain the ID, Case Number, Date, Block, Primary Type, Description, etc. Will keep processing the data set with other data sets.



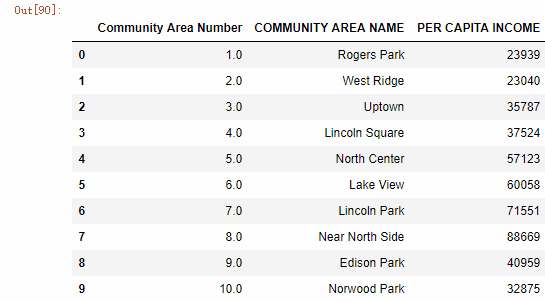
* + Weather data set which contain the Date and the weather type (Will keep processing to change WT01, WT02 … WT20 into fog, cloud, rain, snow, Hail and sunny)



* + The traffic data set after first preprocessing. (Will keep going to preprocess the Volume after combine it with the Community Area)



* Income-Community Area data set (Will keep preprocessing the Income into three level (High, Medium, Low) )



1. Potential Challenge
   * Extract the information we need from U.S. Hourly Precipitation Data database of National Oceanic and Atmospheric Administration, Department of Commerce because their database has so many information and the format is complicated. So we need more time to understand the structure and get the information we need.
   * Overfitting may happen during classification. Since we need to do a lot of classification in this project, overfitting may happen. In order to avoid it, we need to control the model complexity when we apply the model to the data.
2. Techniques
   * To find the best split of income range. To decide which income range should be classified as high income, medium income and low income, we can use decision tree classification. We can try different split points then build the decision tree then calculate total entropy. The pair of split points way that has the smallest entropy is the one we need.
   * To find the best split points of traffic passing volume range which can split the traffic passing volume into three different level -- high, medium and low, we can also use decision tree classification and calculate the GINI index of different split points and find the best split points which has the minimum value of GINI index. Then, we can split the original passing traffic volume into high, medium and low.
3. Tasks Completed

* Create database about crimes
* Create database about traffic situation in each day
* Create database about census data in each area

1. Tasks to be Completed
   * Create database about weather situation in each day
   * Store the data into SQL database by using sqlite 3
   * Merge weather condition data to crime database
   * Merge census data to crime database
   * Merge traffic situation data to crime database
2. Project Timeline

### Phase 1 (2/12 - 3/18)

### Phase 2 (3/19 - 4/29)

#### Week 1 (3/19 - 3/25)

* + - Preprocess the weather data sets

#### Week 2 - 3 (3/26 - 4/8)

* + - Combine all sub datasets into the final dataset
    - Calculate the crime rate base on the final dataset

#### Week 4 - 6 (4/9 - 4/29)

* + - Model building and evaluation base on the final dataset
    - Use different techniques to train the model
    - Calculate the prediction accuracy for future prediction

1. Role of Team Members

#### Chenshu Xu:

1. Keep preprocess the crime data, remove the missing value (which doesn’t have Date or Block or Community Area or Primary Type )
2. Separate the crime data (after preprocess) into sub-dataframes with sub-regions.
3. Combine those sub dataframes of the crime data and the auxiliary data into a final dataset with different subregions.
4. Use final data set to build predictive modeling.
5. Apply different techniques to the predictive modeling to train the dataset. Calculate the prediction accuracy.

#### Hongyu Yan:

1. Preprocess the weather datasets from 2001 to 2018, remove the datas without date and modify them from the original data formula (WT01, WT02, etc.) into new formula (fog, cloud, rain, snow and sunny)
2. Separate the auxiliary data sets into sub dataframes with subregions.
3. Combine those sub dataframes of the crime data and the auxiliary data into a final data set with different subregions.
4. Remove the missing values and noise values from final data set and then calculate the crime rate.
5. Apply different techniques to the predictive modeling to train the dataset. Calculate the prediction accuracy.