ETL-Project - Enrique

## Project Description

For this project, we will download crime statistics from three North Carolina cities: Raleigh, Durham and Cary. Data cleanup and transformation will be accomplished by using MySQL Workbench and Python using dependencies: Pandas, SQLalchemy.

Crime statistics are easily located, however, every jurisdiction reports those statistics in a unquie format. We hope to find relationships, trends, contrasts by making the data uniform across these different cities.

## Data Sources and \*\*E\*\*xtract

We started by extracing the Crime reports from [data.world] (https://data.world/) in .csv format.

\* Raleigh, NC:

https://data.world/mschnars/raleigh-crime-incidents/workspace/file?filename=RaleighPoliceIncidents.csv

\* Durham, NC:

https://data.world/durhamnc/durham-police-crime-reports/workspace/file?filename=city-of-durham-police-crime-reports\_4.csv

\* Cary, NC:

https://data.world/townofcary/crime-mapping/workspace/file?filename=crime-mapping\_3.csv

The downloaded .csv files were put into different Pandas Dataframes for review and cleaning.

## Cleaning and \*\*T\*\*ransform:

For consistency, we settled on keeping following categories in each dataframe:

1. Date (Crime Reported - day of the week, year/month/day)

2. Time of Crime

3. Crime Type

\* Each jurisdiction labels similar crimes differently i.e. Durham has Nine different Larceny categories.

\* Crime categorized tables were created for each of the cities to compare like Crime Types

\* The Crime Categories choosen were Assult, Murder, Stolen, and Other.

4. City Name

Data Transformation included:

\* Cleaning date format for consistency,

\* Date/time conversions,

\* Key restructuring,

\* Joining crime categorized tables with City Crime statistics,

\* Filtering to keep columns we needed and

\* Integrating heading names and crime caegories.

\* Concatinated city dataframes into one master dataframe

## Collection and \*\*L\*\*oad