ETL Project Technical Report by David Wang, Susi Ng

Introduction

For this ETL project, we selected to obtain crime data in the cities in California and also looking at certain factors of the local demographic such as education level, median income, and unemployment rate.

Description of Process

Extract

We extract our data from two online sources which showed details at City Police Department level:

Demographics: <https://openjustice.doj.ca.gov/data>

Crime Information: <https://www.ucrdatatool.gov/index.cfm>

Both sources have downloads available in CSV format. We use Pandas to import into dataframe.

Transform

We begin with determining which data columns we want to put into new data frame and rename. We use Pandas for this cleaning and analyzing process. Next we split the columns Agency to obtain just the city name, as we’d like to be able to have a common field between the two datasets as primary key to join later. We also removed rows that are: 1) not representing individual city data, 2) not year 2014. Later we reset the index for the new data frame.

Load

Using MySQL Workbench, we created a new SQL database and created two tables for uploading the data we just transformed from Pandas. We utilize SQLAlchemy and PYMYSQL to make the local connection to database for running the query. We are able to load the data to database successfully and merge/ make a join query.

Conclusion

The ETL process enable us to combine information from various sources into datasets such as datasets on crime and the datasets on demographic which we can analyzed further based on observable relationship.