**ETL Project**

**August 22, 2019**

**Task Summary:**

Select two sources of data and perform data extraction, transformation and load the data into a new database.

**Team Members:** Sasha, Veena and William

**EXTRACTION:**

I. Extracted .csv files from data sources.

1. Read .csv data files into an Excel Workbook.

2. Built pgAdmin tables and columns from .csv files.

3. Used Panadas to connect to the database.

II. Datasets:

1. U.S State vs. Ethnicity rate (e.g White, Black, Hispanic, Asian, Two or More Races)
   1. Source: Kaiser Family Foundation estimates based on the Census Bureau's American Community Survey, 2008-2017.
   2. Data format: Comma Separated Values file (.csv)
2. U.S State vs. Unemployment Rate (2010)
   1. Source: Bureau of Labor Statistics (<https://data.bls.gov/lausmap/showMap.jsp>)
   2. Data format: Comma Separated Values file (.csv)
3. U.S State vs. Crime Rate (e.g. Violent, Murder, Forcible rape, Robbery, Aggravate assault, Total)
   1. Source: <https://www.infoplease.com/us/crime/crime-rate-state-2010>
   2. Data format: Comma Separated Values file (.csv)
4. U.S State vs. Education Attainment (e.g. Percent high school graduate or higher, Percent bachelor's degree or higher)
   1. Source: U.S. Census Bureau, American Community
   2. Data format: Comma Separated Values file (.csv)
5. U.S State vs. Population (2010)
   1. Source: U.S. Census Bureau, American Community
   2. Data format: Comma Separated Values file (.csv)

**TRANSFORMATION**

We used Panadas as our primary tool to clean the data. Each csv file was read and created as data frame. Columns were renamed and dropped/added accordingly. All data frames were joined to create a final ‘ETL\_data’ dataframe.

**LOAD**

The final data frame ‘ETL\_data’ was transferred to the ‘etl\_project’ SQL database (created in PgAdmin). This was done by creating an engine and connecting string.

We tested our new database by joining two table and ran a query to determine if the data was loaded without errors. Our objective was to find out if we could correlate the data to look at population trends, education and crime.